



Memorias Cache

Arq. de Computadores

Santiago González Tortosa



Parte I

Introducción a Memorias Cache

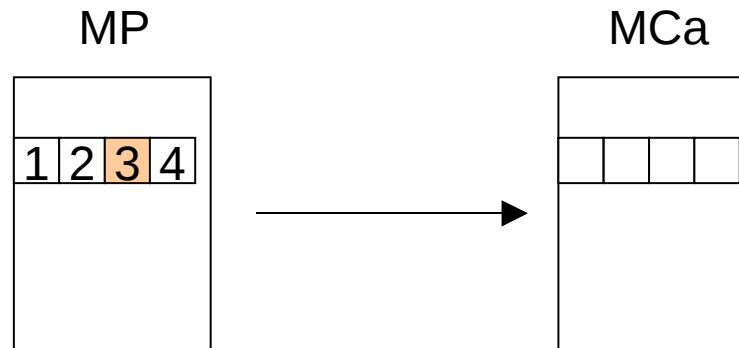
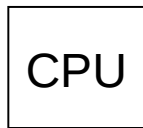


Introducción

- Para que sirve
 - Reducción de tiempo en el acceso de información (no acceder a MP)
- Tiempo de acceso reducido
- Control de Información
 - Política de lectura: Para un buen uso de Mca, se necesita tener actualizada la información en la misma (coherencia). ¿Cómo se lleva la info a Mca?
 - Política de escritura: Cuando el *user* quiere modificar info, ¿cómo se escribe?
- **OJO:** Reemplazo de Mca únicamente en asociativas y asociativas por conjuntos

Política de Lectura

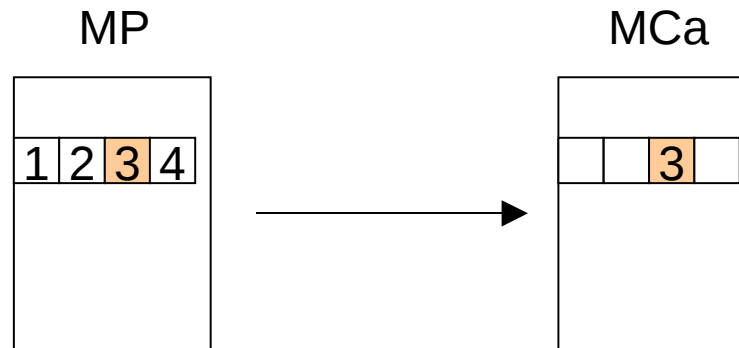
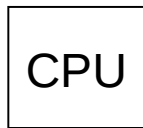
■ OOF



■ Early Start

Política de Lectura

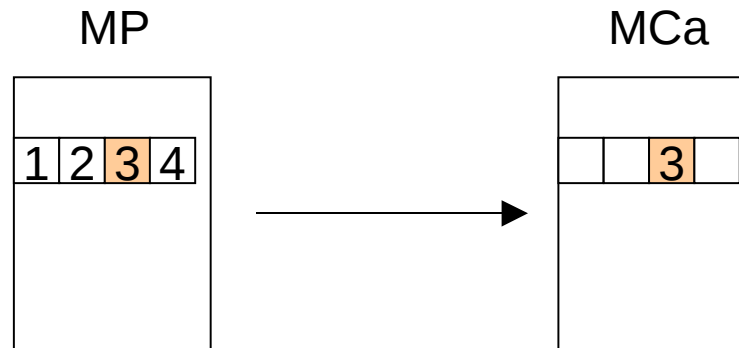
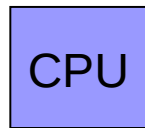
■ OOF



■ Early Start

Política de Lectura

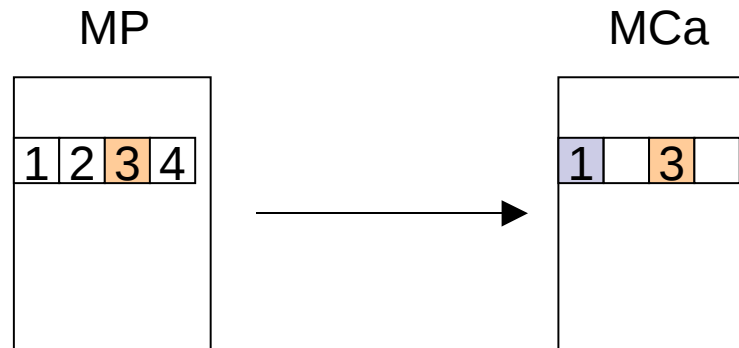
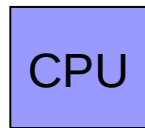
■ OOF



■ Early Start

Política de Lectura

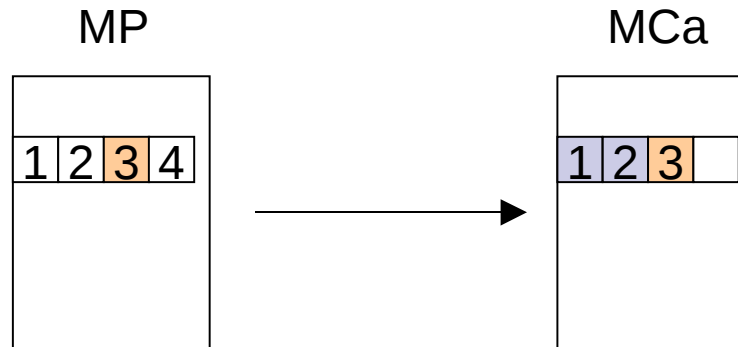
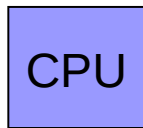
■ OOF



■ Early Start

Política de Lectura

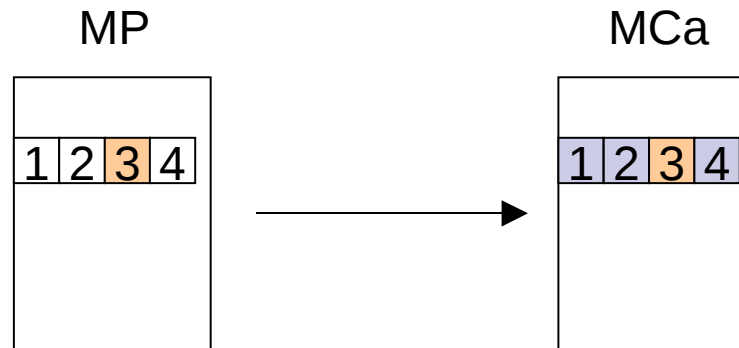
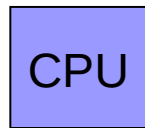
■ OOF



■ Early Start

Política de Lectura

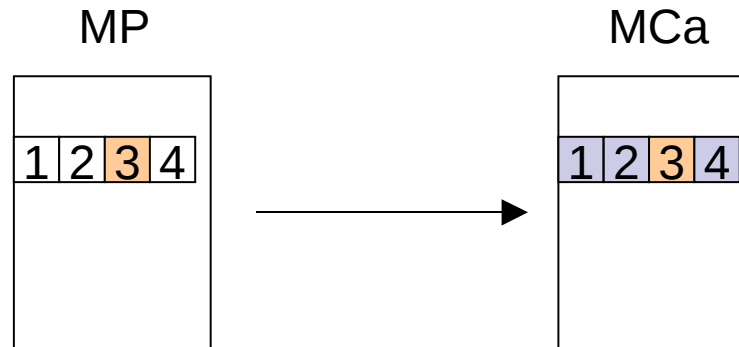
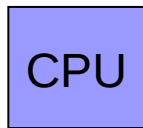
- OOF



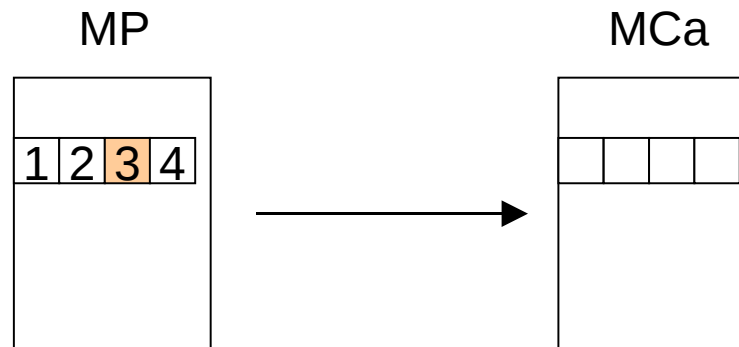
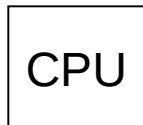
- Early Start

Política de Lectura

■ OOF

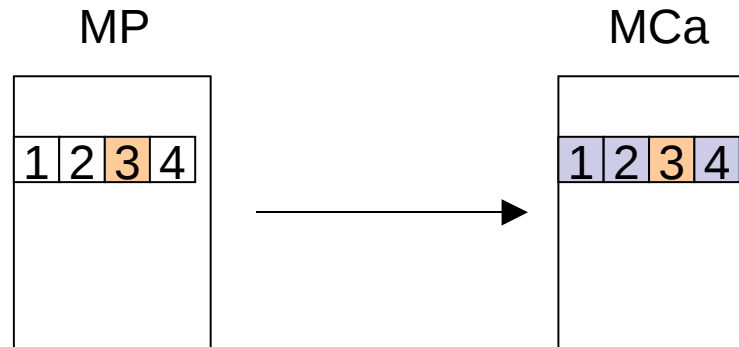
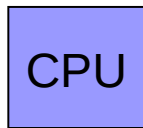


■ Early Start

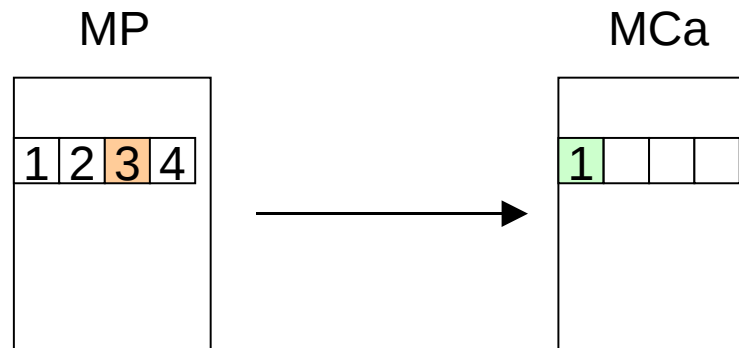
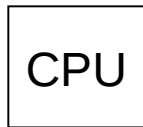


Política de Lectura

■ OOF

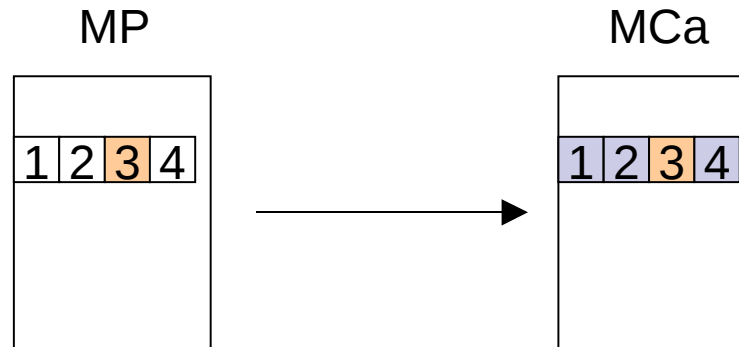
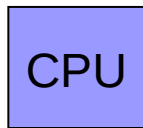


■ Early Start

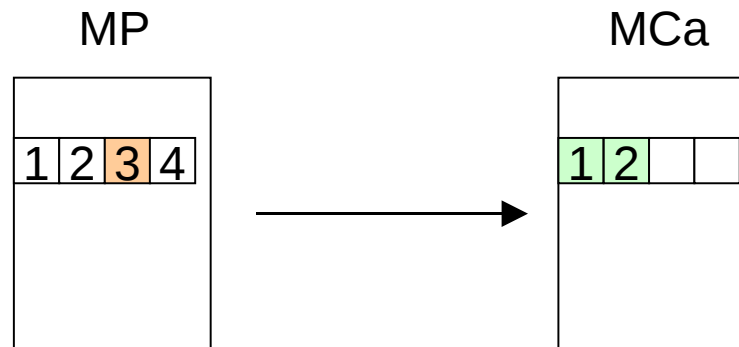
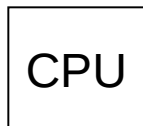


Política de Lectura

■ OOF

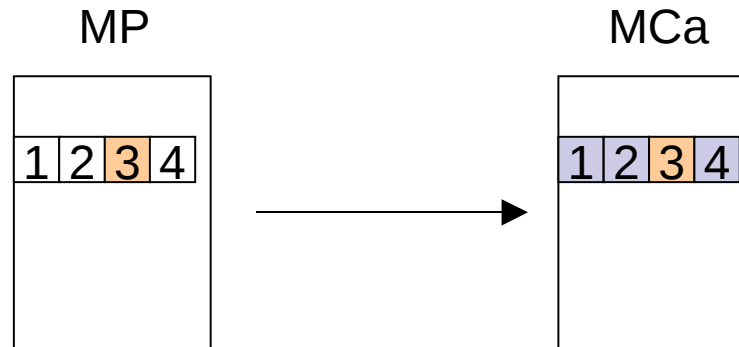
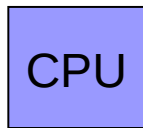


■ Early Start

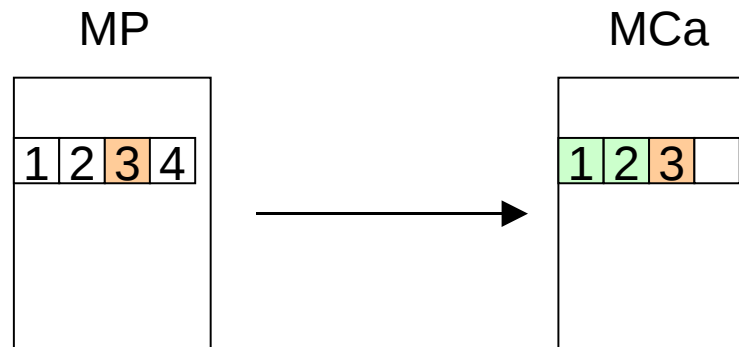
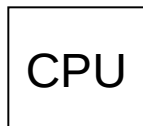


Política de Lectura

■ OOF

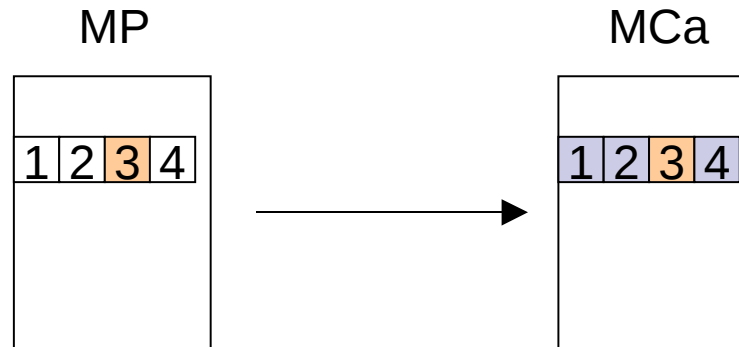
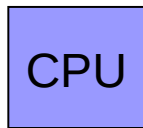


■ Early Start

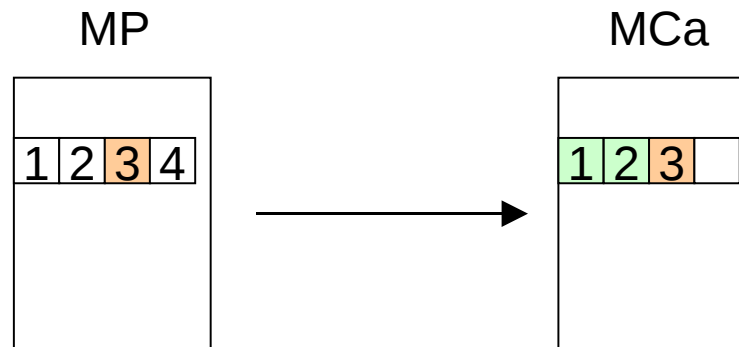
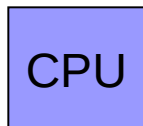


Política de Lectura

■ OOF

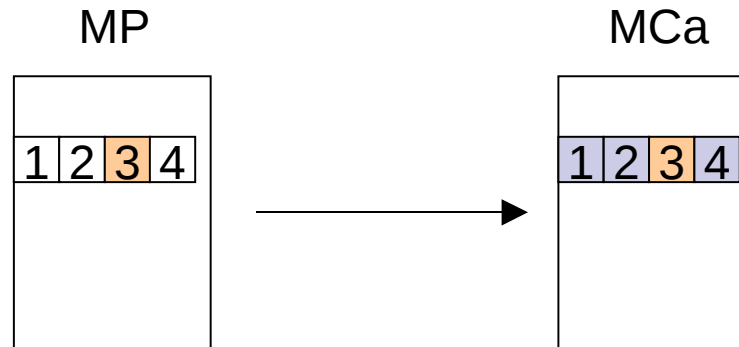
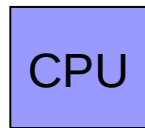


■ Early Start

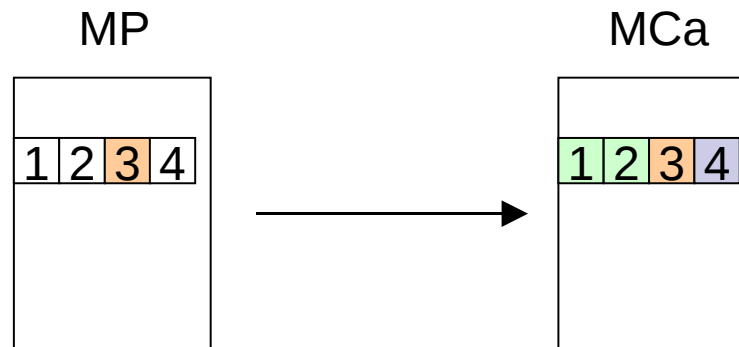
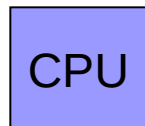


Política de Lectura

■ OOF

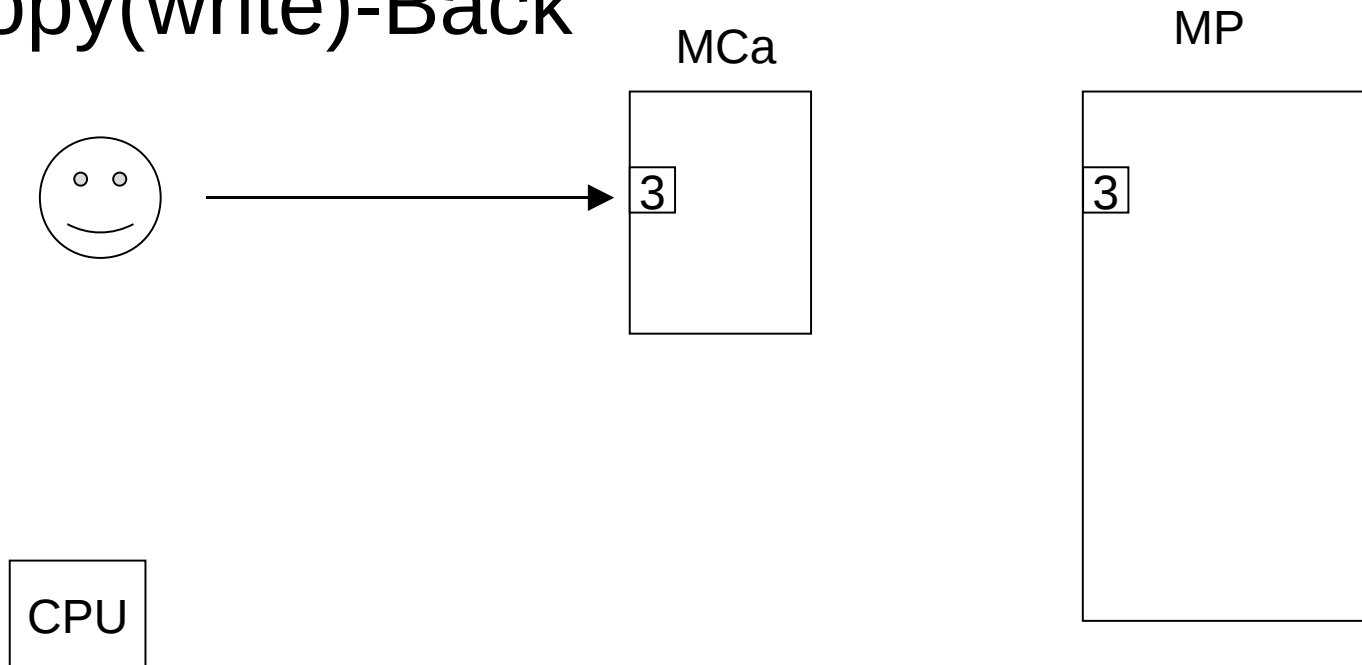


■ Early Start



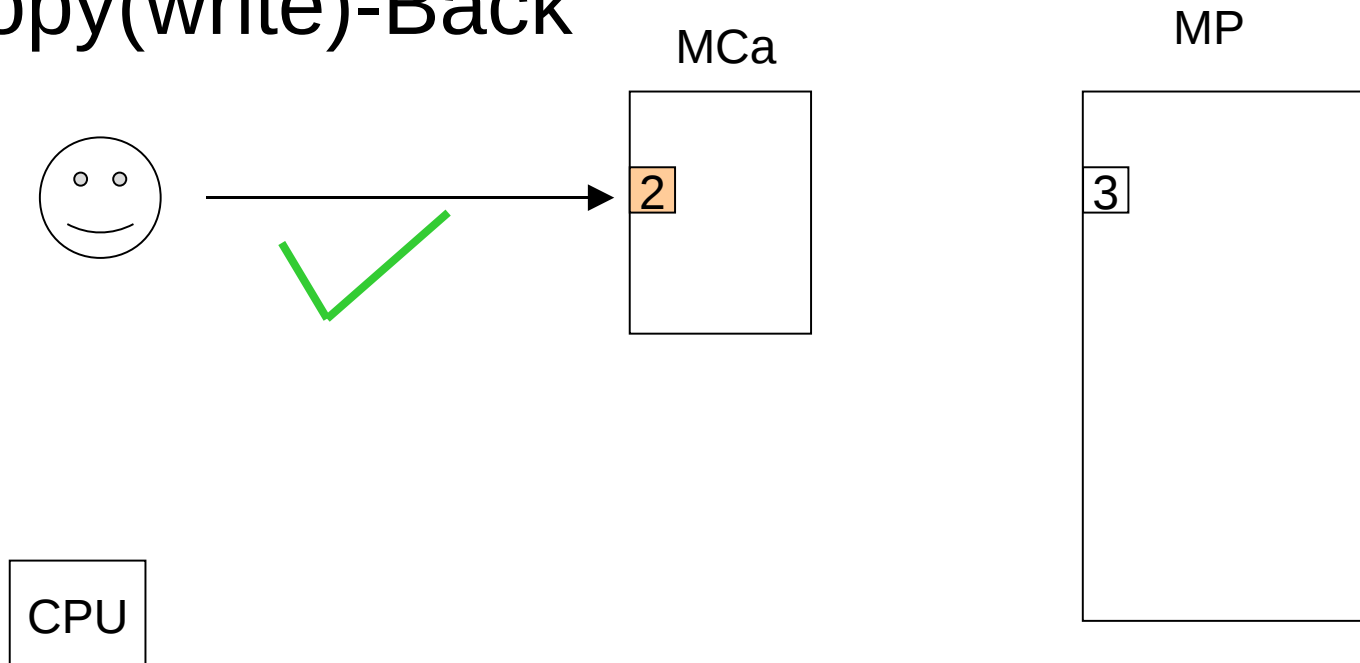
Política de escritura

■ Copy(write)-Back



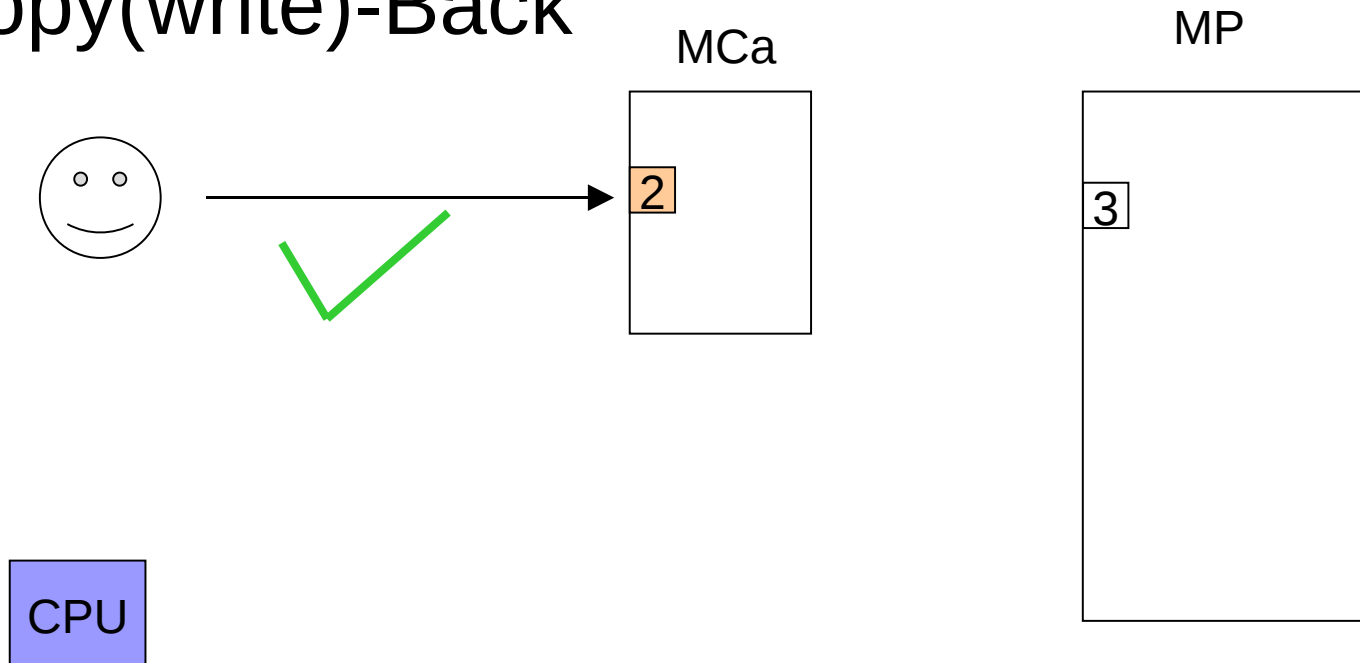
Política de escritura

■ Copy(write)-Back



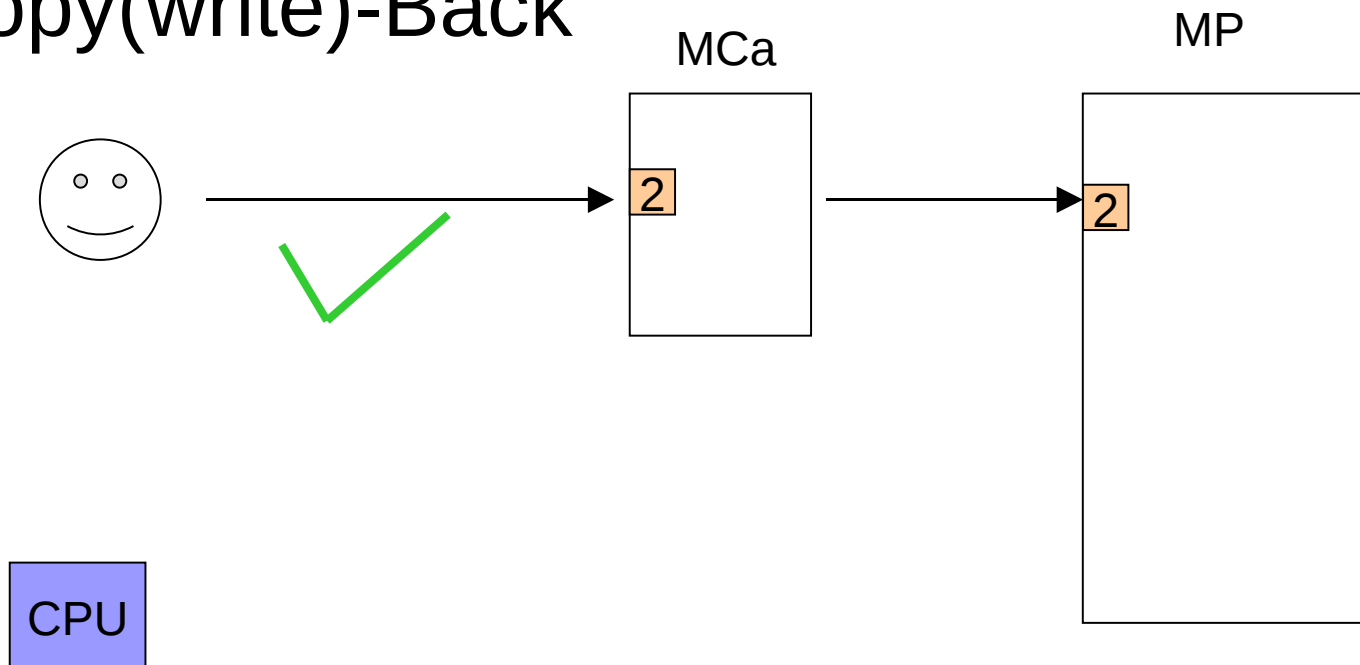
Política de escritura

■ Copy(write)-Back



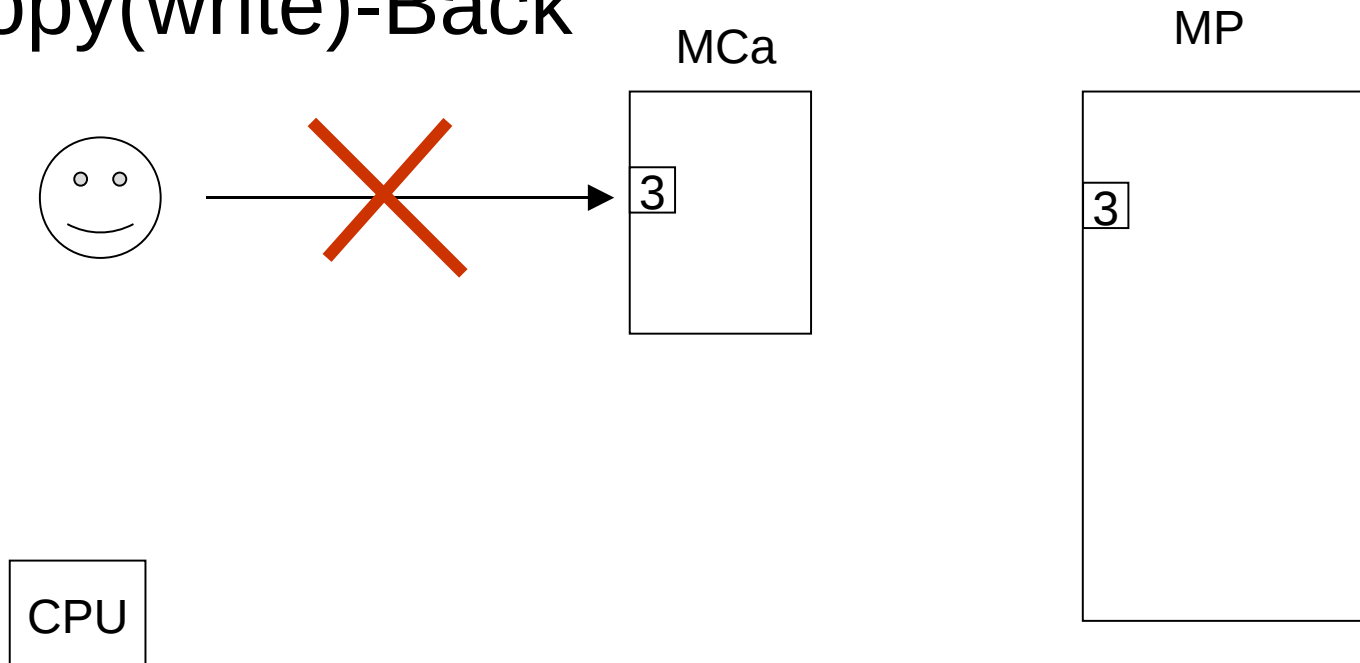
Política de escritura

■ Copy(write)-Back



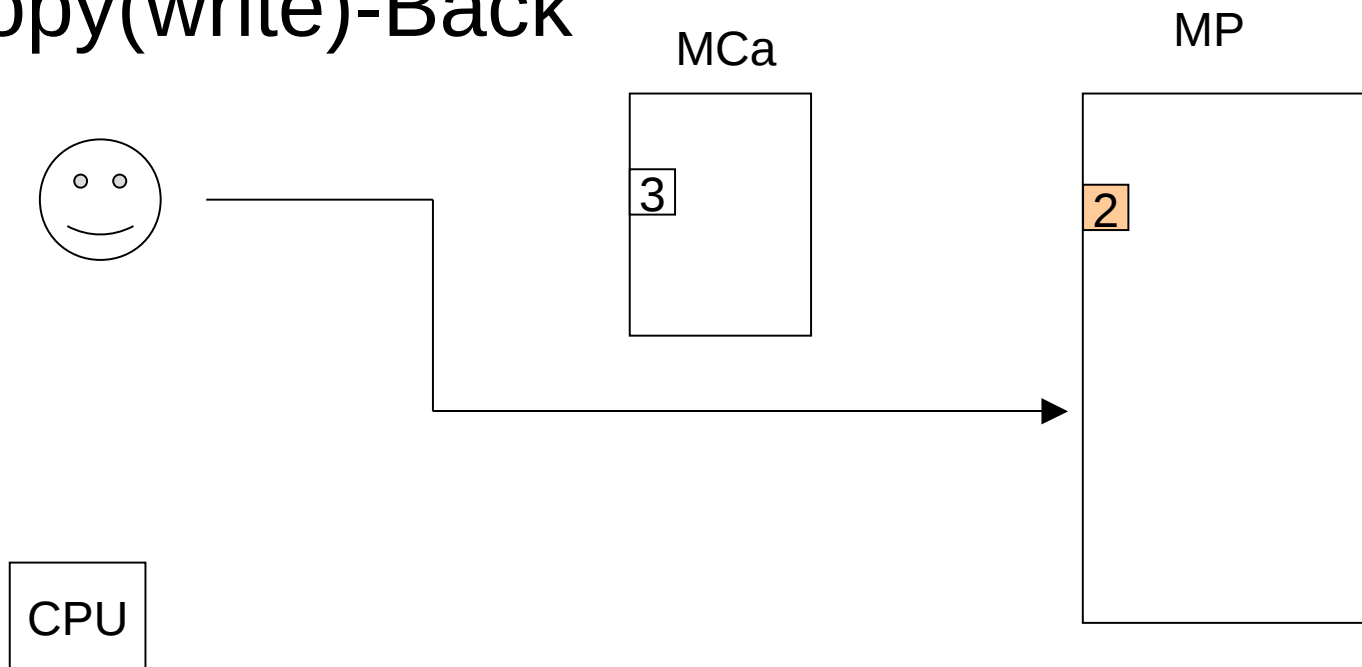
Política de escritura

■ Copy(write)-Back



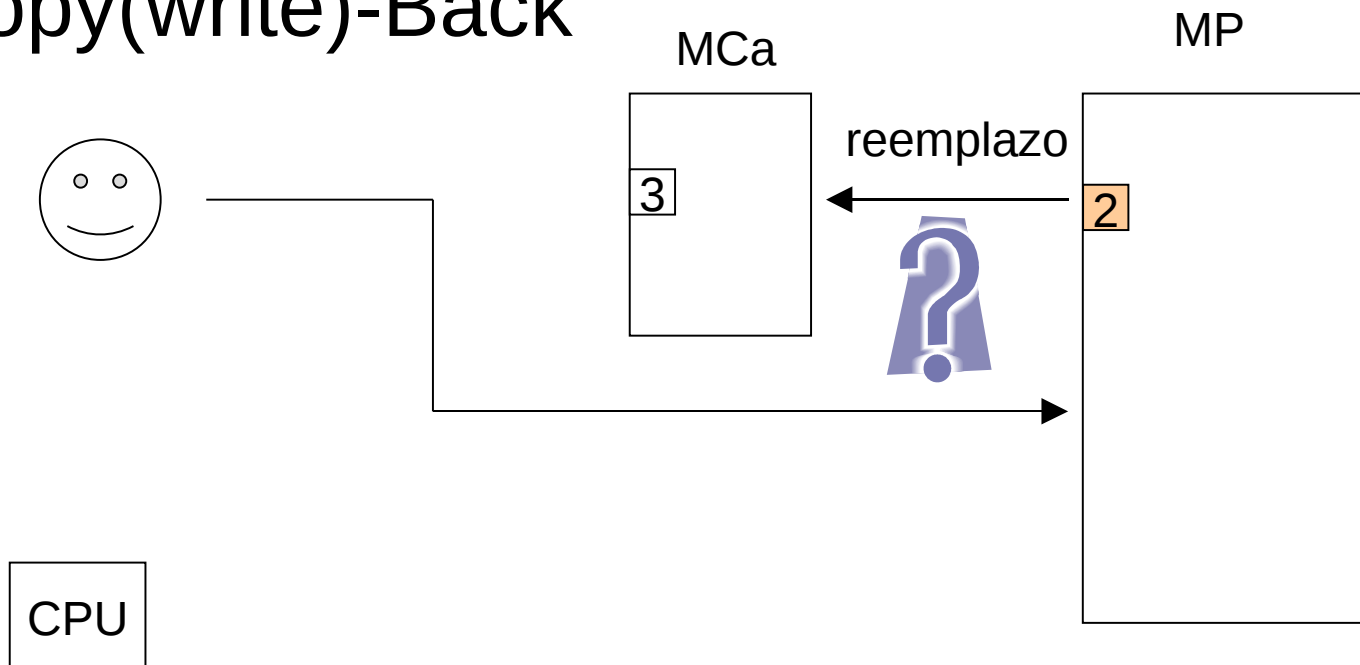
Política de escritura

■ Copy(write)-Back



Política de escritura

■ Copy(write)-Back

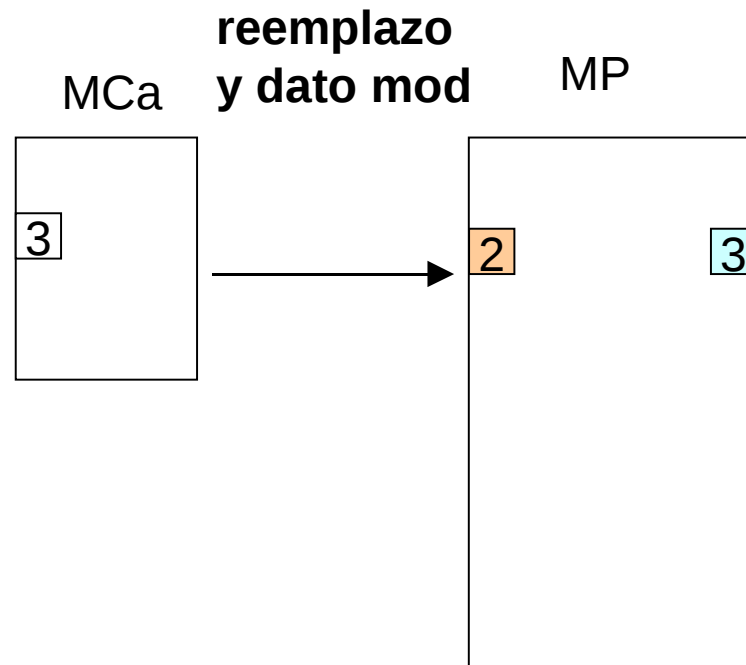


Política de escritura

■ Copy(write)-Back



CPU

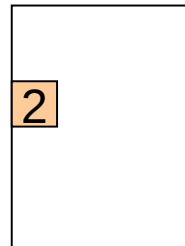


Política de escritura

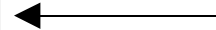
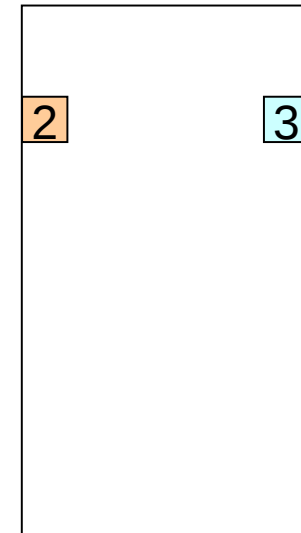
■ Copy(write)-Back



MCa



MP



Segun Politica Lectura

CPU

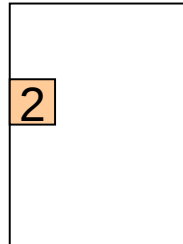


Política de escritura

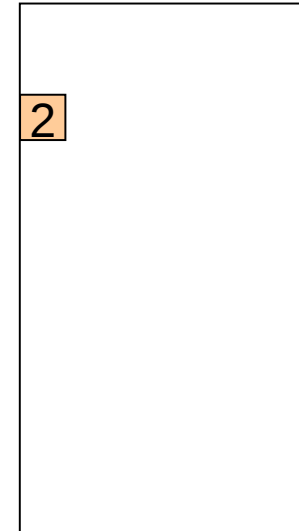
■ Copy(write)-Back



MCa



MP



Segun Politica Lectura

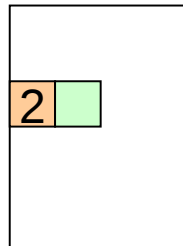
CPU

Política de escritura

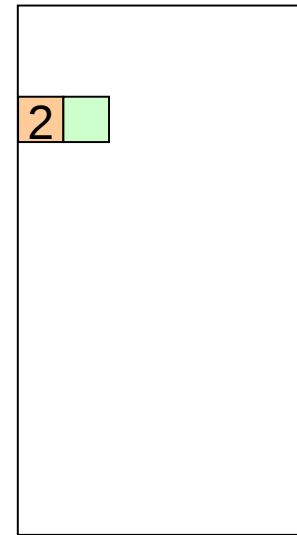
■ Copy(write)-Back



MCa



MP



Segun Politica Lectura

CPU



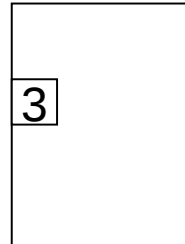
Política de escritura (II)

■ Write-through

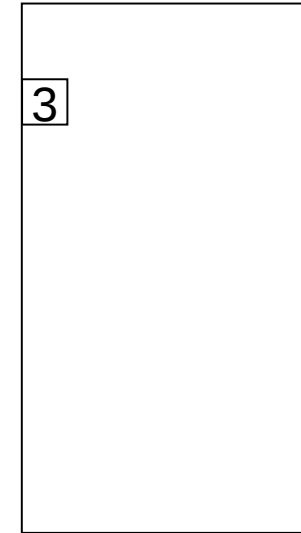


CPU

MCa

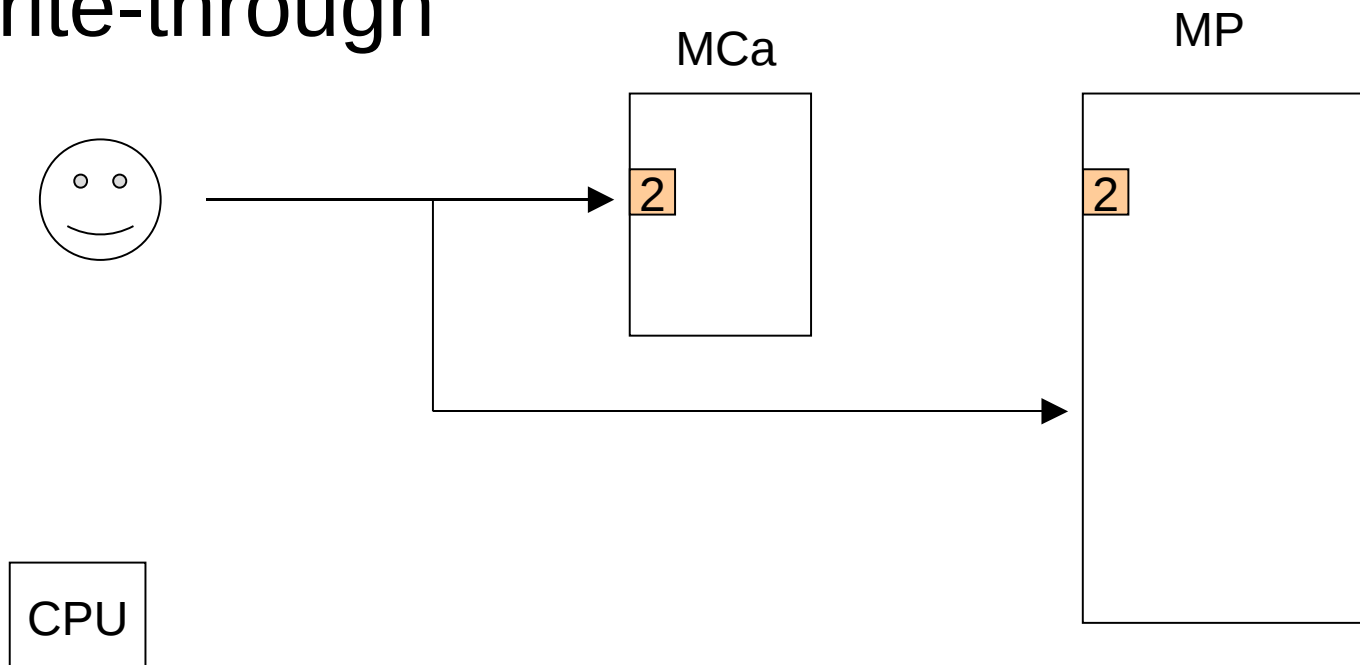


MP



Política de escritura (II)

■ Write-through



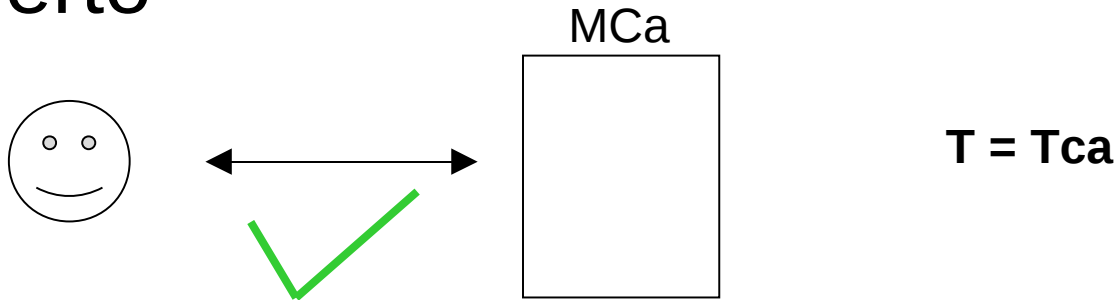


Operaciones MCa

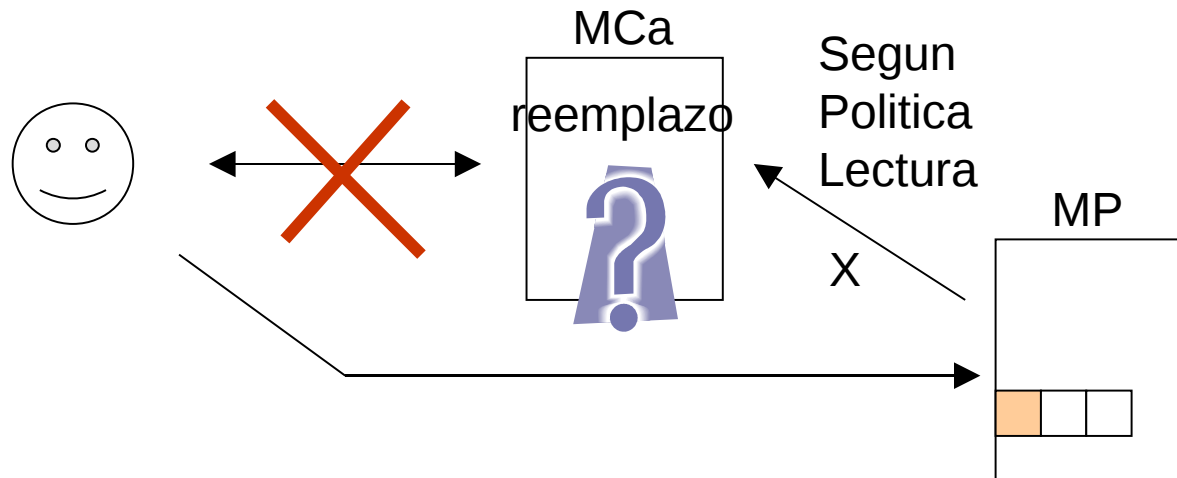
- Lectura
 - ☐ Acierto
 - ☐ Fallo
- Escritura (Política de Escritura)
 - ☐ Acierto
 - ☐ Fallo
- Debemos tener en cuenta políticas

Lectura MCa

■ Acierto

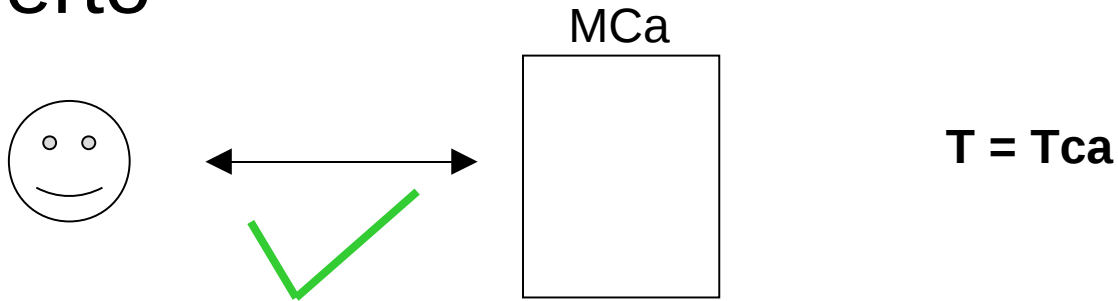


■ Fallo

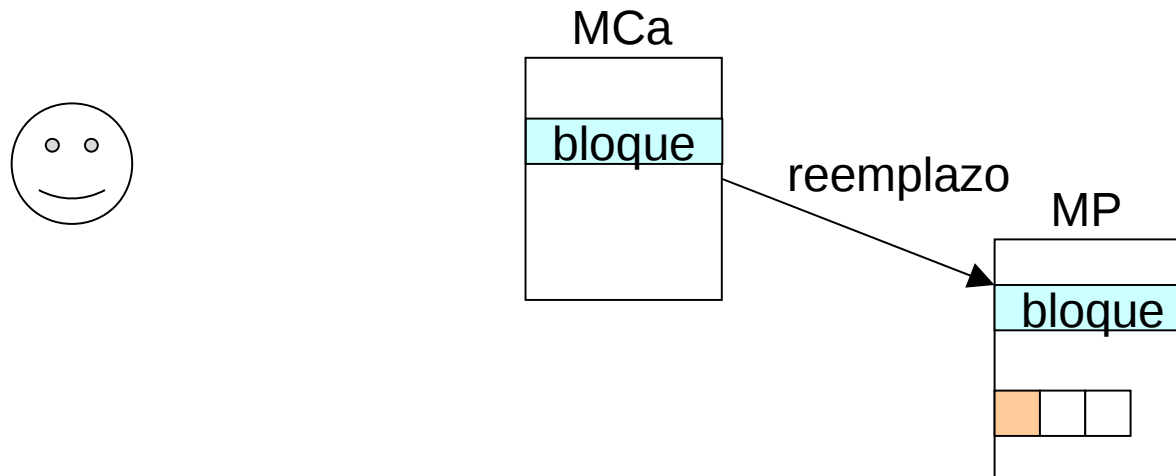


Lectura MCa

■ Acierto

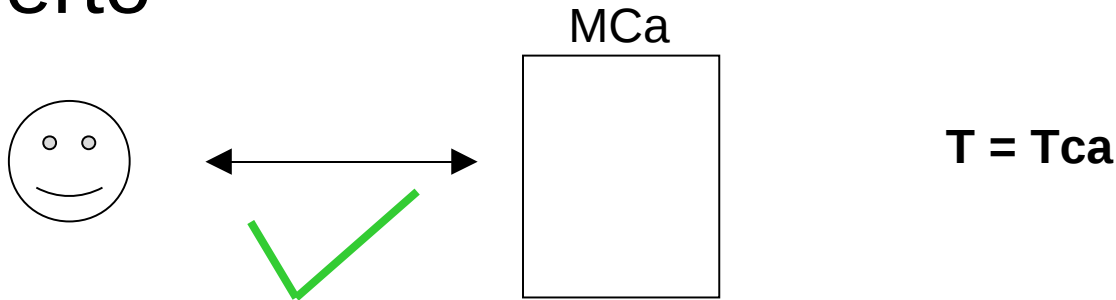


■ Fallo

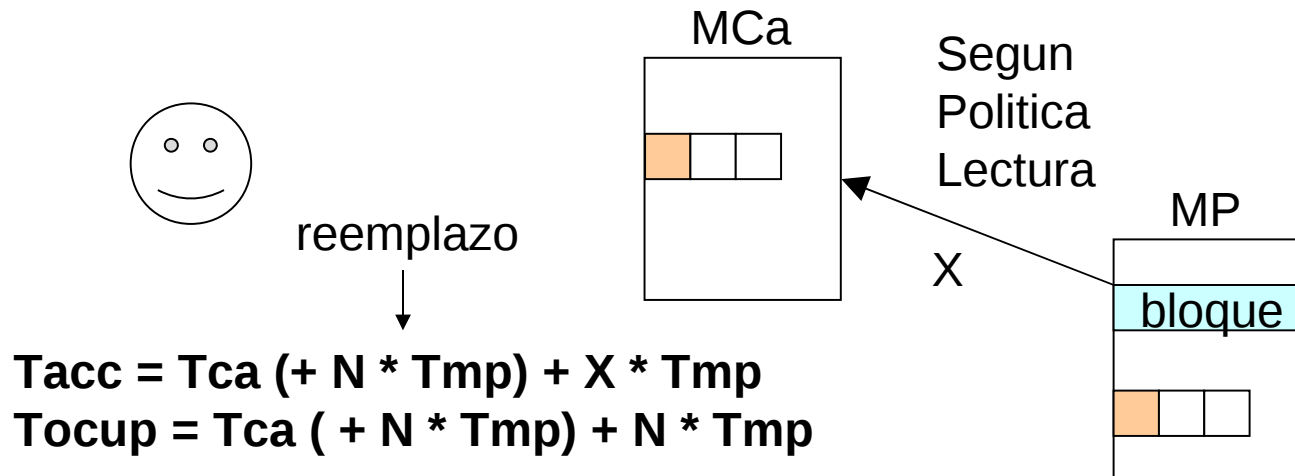


Lectura MCa

■ Acierto



■ Fallo



Posibles Derivaciones

- Tiempo medio de Acceso: probabilidades de Acierto/Fallo
 - $T = Hr * T_{acierto} + (1-Hr) * T_{fallo}$
- Tiempo medio de Acceso con Probabilidades de Lectura/Escritura
 - $P_{lec} * T_{lec} + P_{esc} * T_{esc}$
 - T_{lec} calculado con probabilidades acierto/fallo
 - T_{esc} calculado con probabilidades acierto/fallo

Posibles Derivaciones (II)

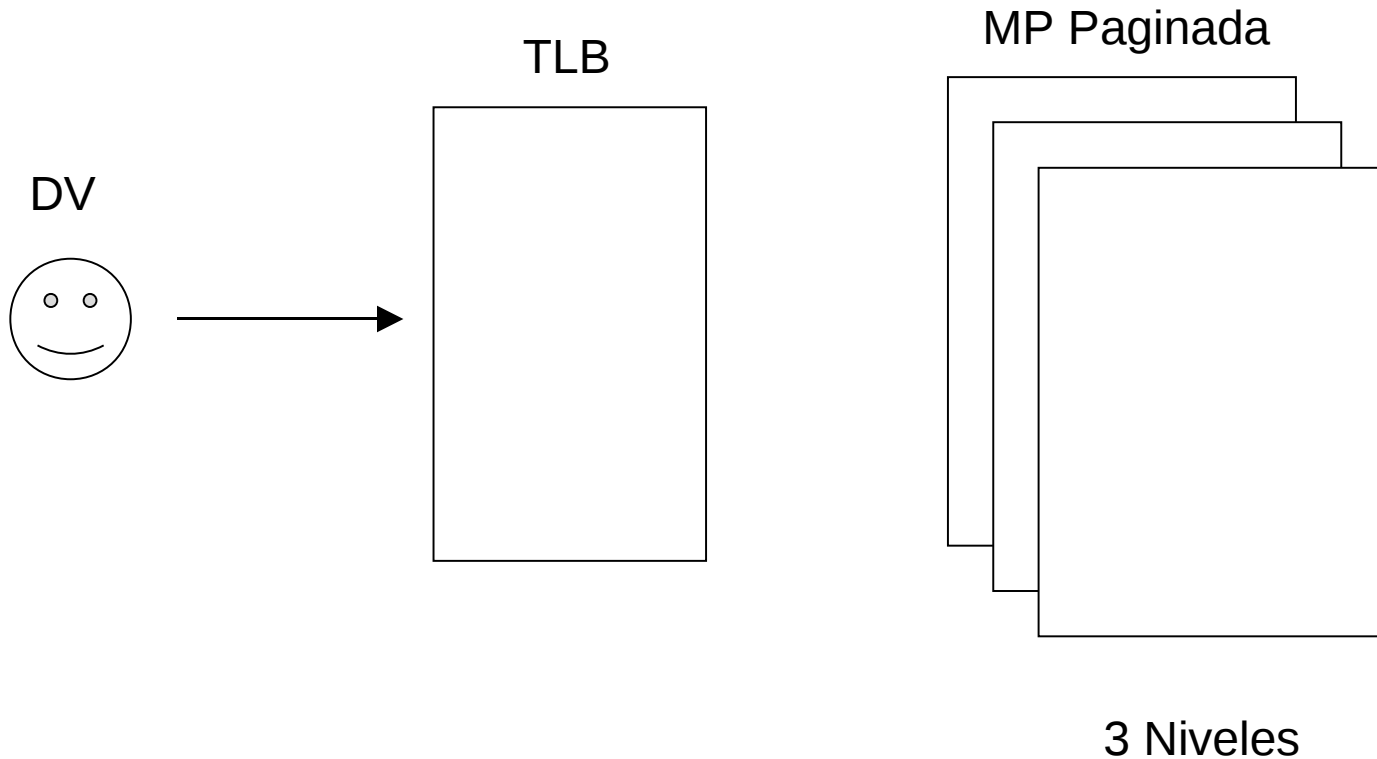
- Tiempo máximo de acceso
 - Igual que tiempo medio de acceso, pero **sin tener en cuenta la posibilidad de acierto.**
- Tiempo mínimo de acceso
 - Igual que tiempo medio de acceso, pero **sin tener en cuenta la posibilidad de fallo.**
- Tiempo de ocupación
 - Igual que el tiempo medio de acceso, pero **teniendo en cuenta la lectura y escritura de TODAS las palabras del bloque** (tiempo en que la CPU realiza otras operaciones)



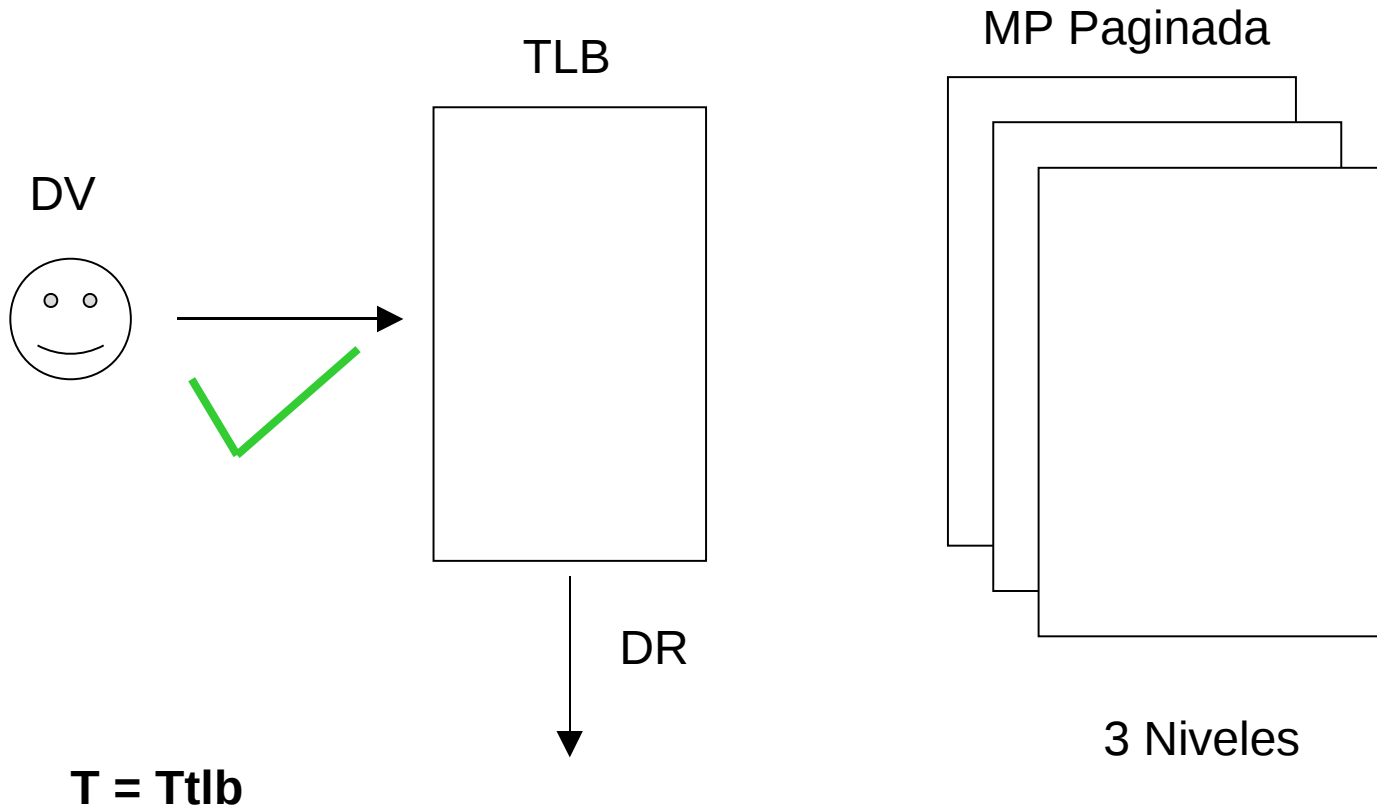
Parte II

Uso de Memoria Virtual

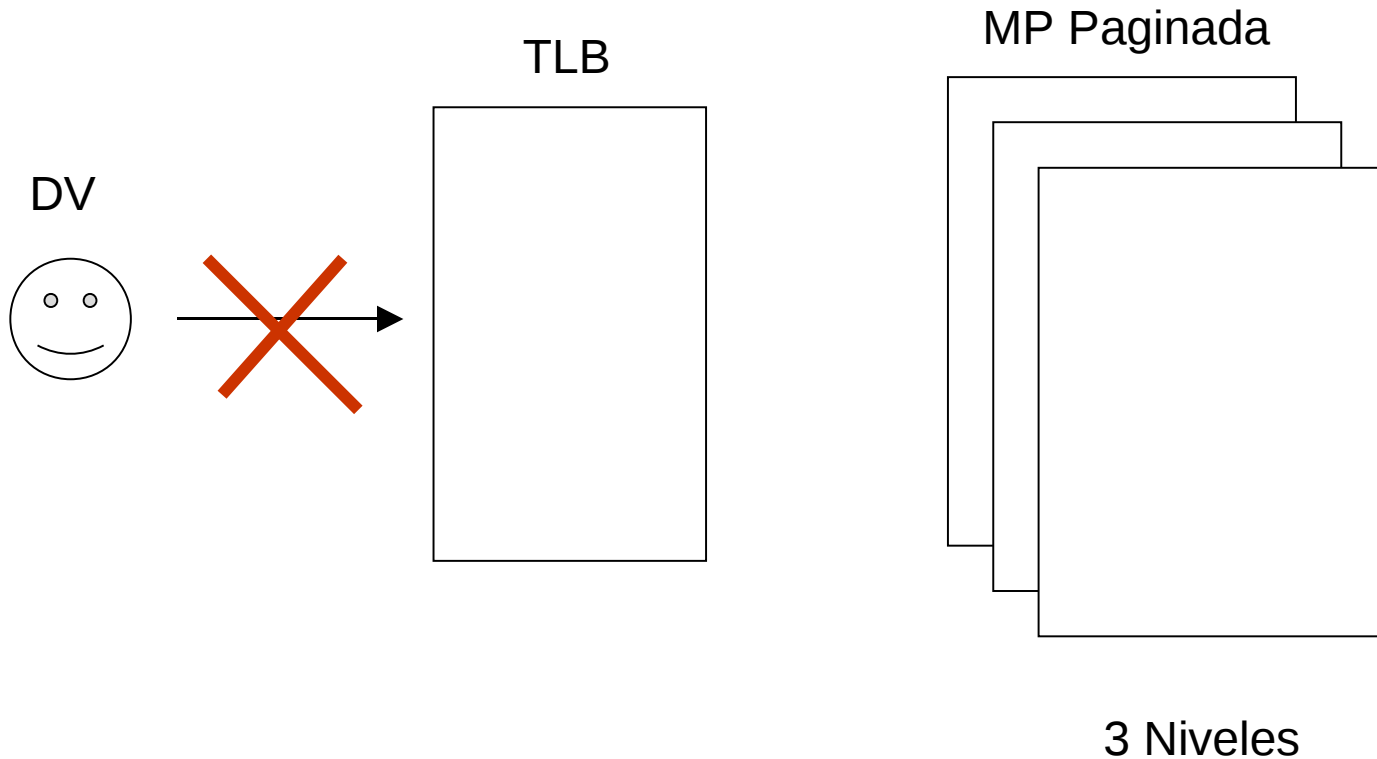
Fase de Traducción



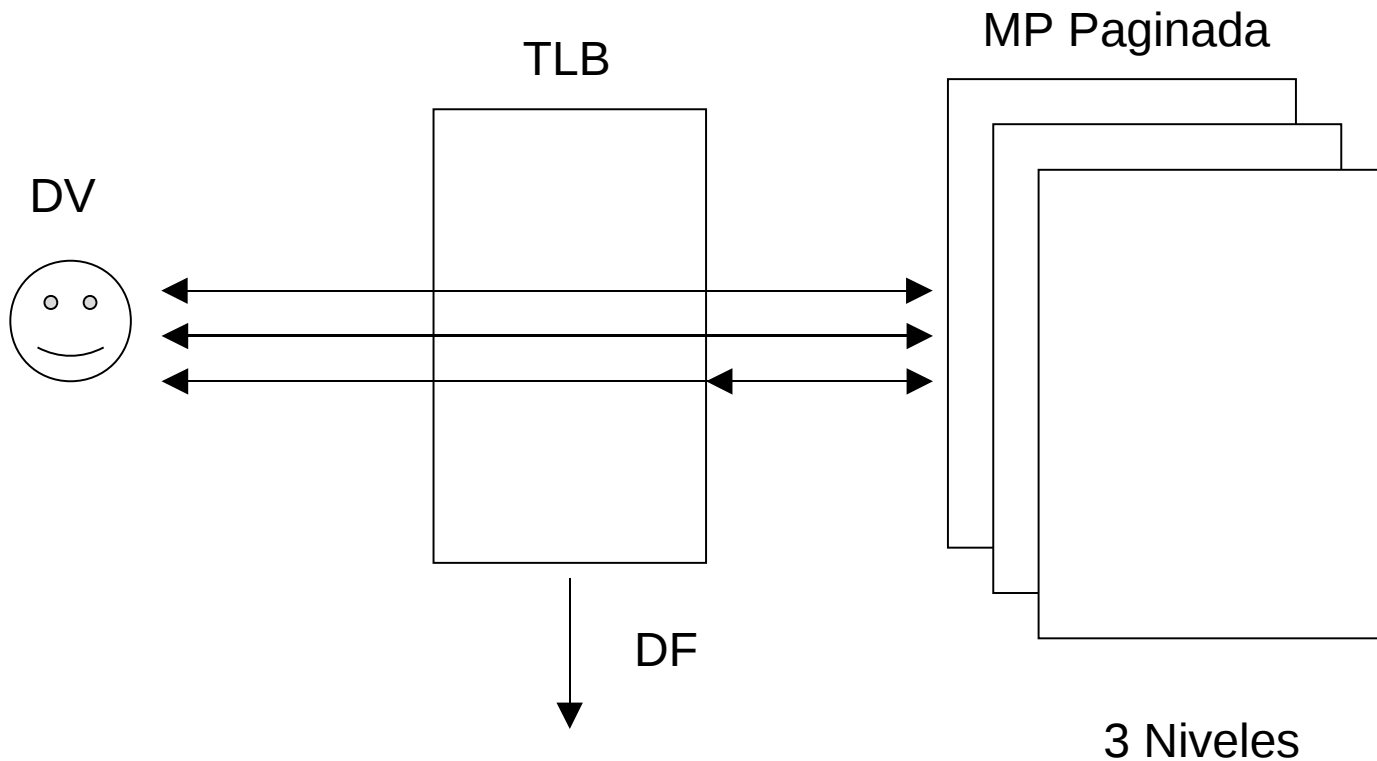
Fase de Traducción



Fase de Traducción



Fase de Traducción



$$T = T_{tlb} + 3 * T_{mp}$$

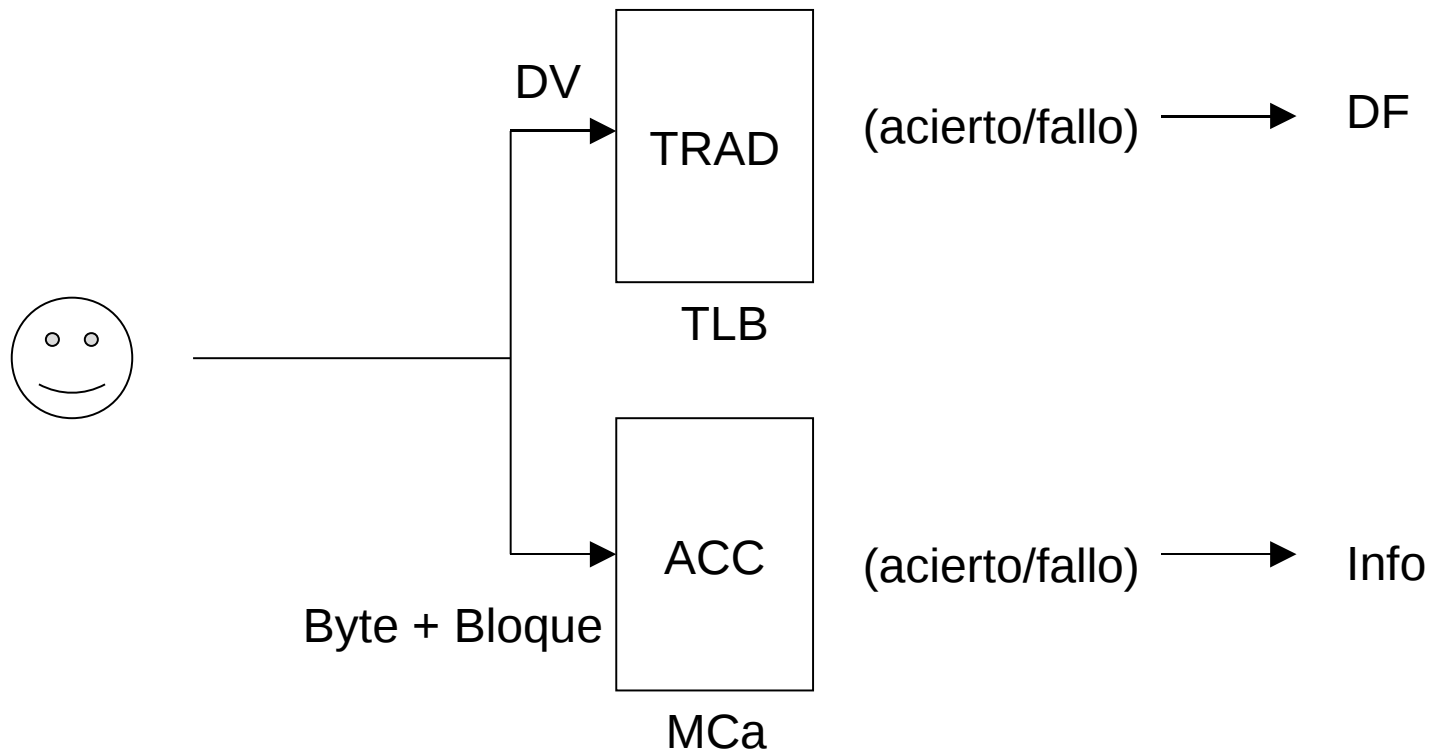
Uso de Memoria Virtual

■ Traducción y Acceso a la Info secuencial



Uso de Memoria Virtual (II)

■ Traducción y Acceso a la Info Simultaneo



Uso de Memoria Virtual (II)

■ Tiempo de Traducción

- $T = H_{rtlb} * T_{tlb} + (1 - H_{rtlb}) * (T_{tlb} + N * T_{mp})$

■ Tiempo de Acceso

- Acierto: $T = T_{ca}$

- Fallo: $T = T_{ca} (+ \dots)$

■ Habiendo Simultaneidad

- $T = \max(T_{trad}, T_{acc}) + (\text{Si fallo } M_{Ca}) \dots$

Uso de Memoria Virtual (III)

■ Diferencias entre secuencial y simultaneo

□ Acierto en Cache

- $T_{sec} = T_{trad} \text{ (aciertos o fallos)} + T_{ca}$
- $T_{simul} = \max(T_{trad}, T_{ca})$

□ Fallo en Cache

- $T_{sec} = T_{trad} \text{ (acierto o fallo)} + T_{acc}$
- $T_{simul} = \max(T_{trad}, T_{ca}) + T_{fallo}$

Uso de Memoria Virtual

- Posibilidades:
 - Traducción
 - Acierto TLB: T_{tlb}
 - Fallo TLB: $T_{tlb} + N * T_{mp}$ (N niveles de pagina)
 - Acceso Información
 - Acierto Mca: T_{ca}
 - Fallo Mca: $T_{ca} (+ T_{fallo_cache})$
- Mezclar posibilidades con derivaciones de tiempos
- **OJO:** Tiempo de Traducción es siempre el mismo en todos los casos (acceso, ocupación)
 - Tpo maximo de acceso: Solo fallos de TLB ($1 - Hr_tlb = 100\%$)
 - Tpo minimo de acceso: Solo aciertos de TLB ($Hr_tlb = 100\%$)



Parte III

Inclusión de nuevas
tecnologías

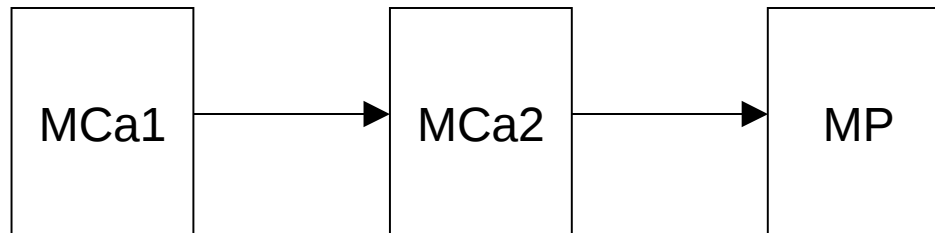


Nuevas Tecnologías

- Memoria Cache de 2º Nivel
 - Stream Buffer
 - Buffer de Escritura
 - Victim Buffer
-
- Posibilidad de mezclar: Traducción + Acceso simple a MCa + Nuevas Tecnologías

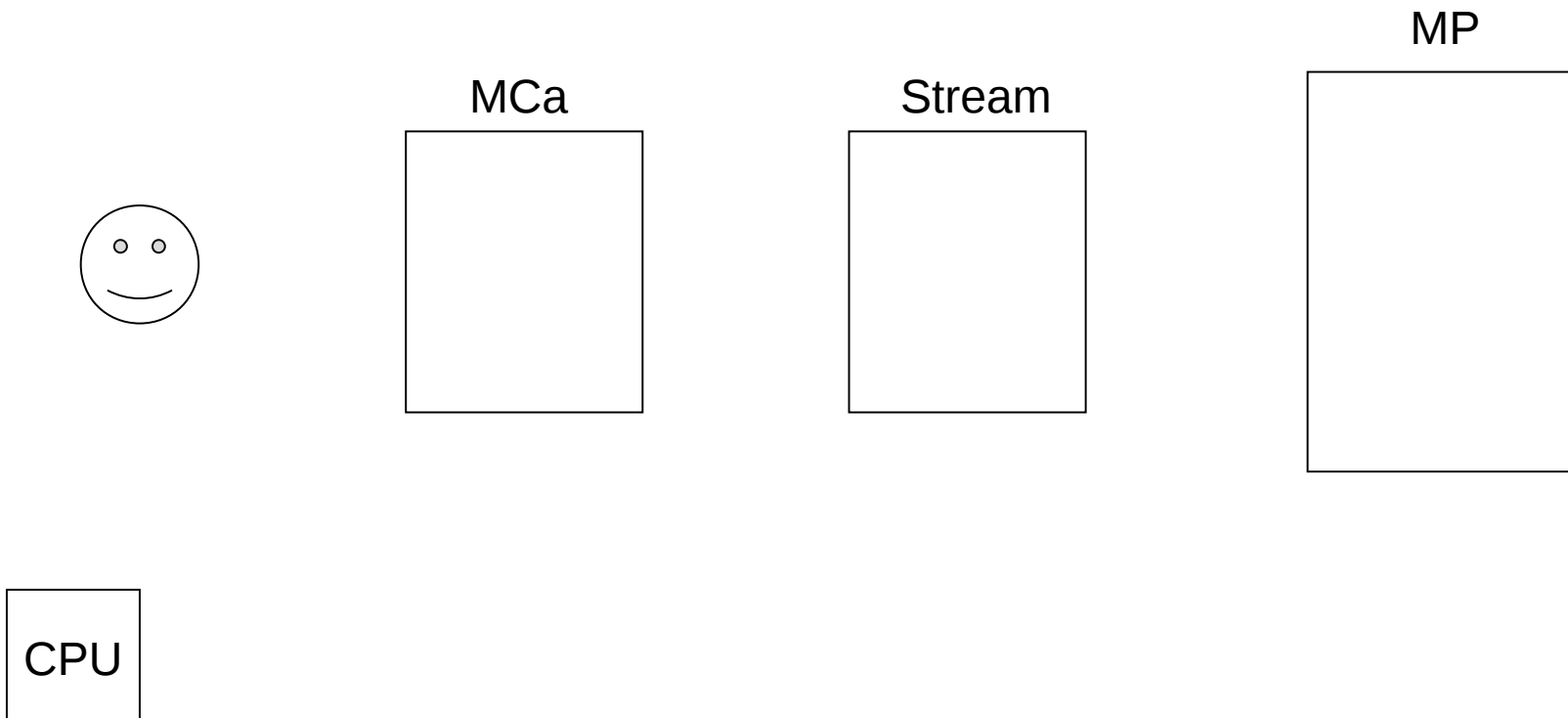
Memoria Cache de 2º Nivel

- Cuando falla el MCa de Nivel 1, se accede al MCa de Nivel 2.
- Políticas de Nivel 1 y Políticas de Nivel 2
- Si falla MCa de Nivel 2, se accede a MP.



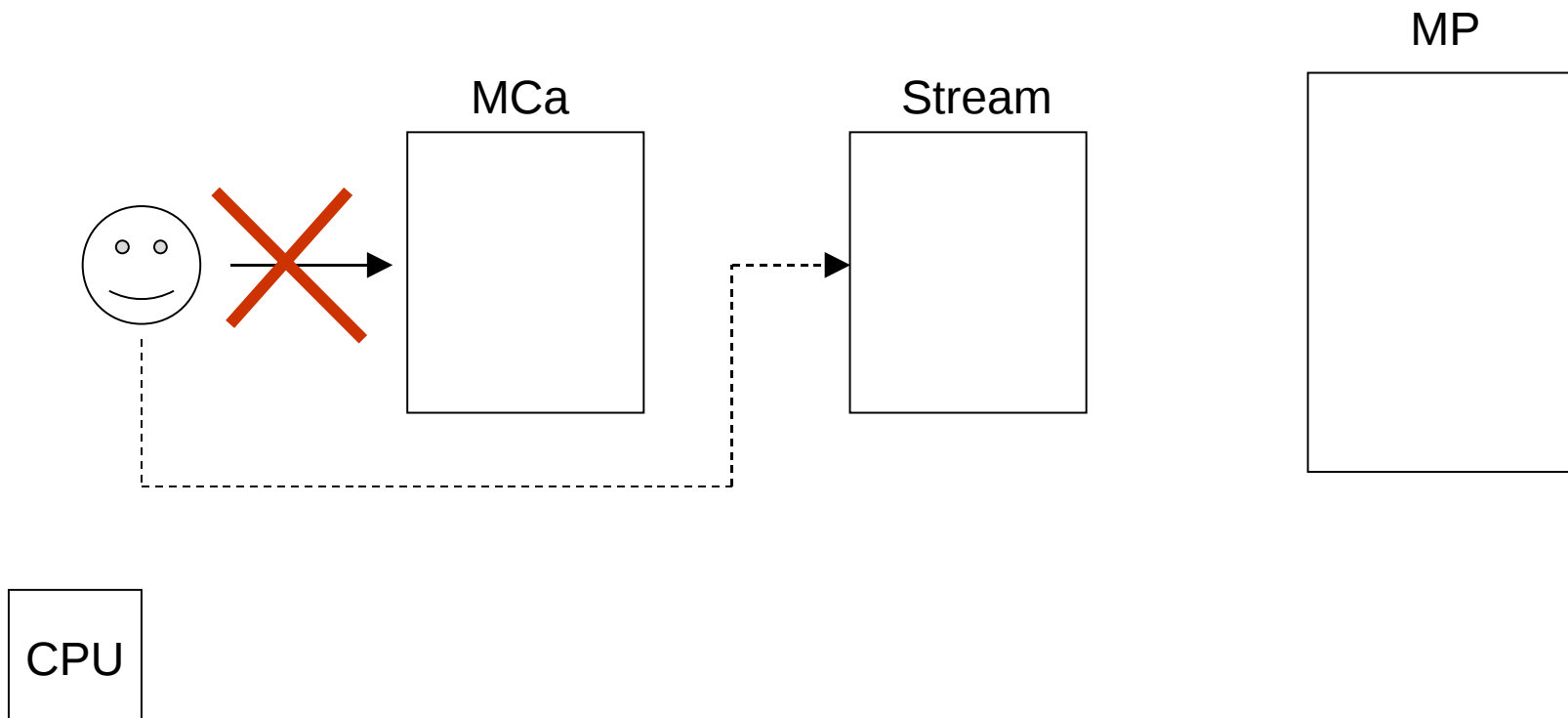
Stream Buffer

■ Anticipaciones



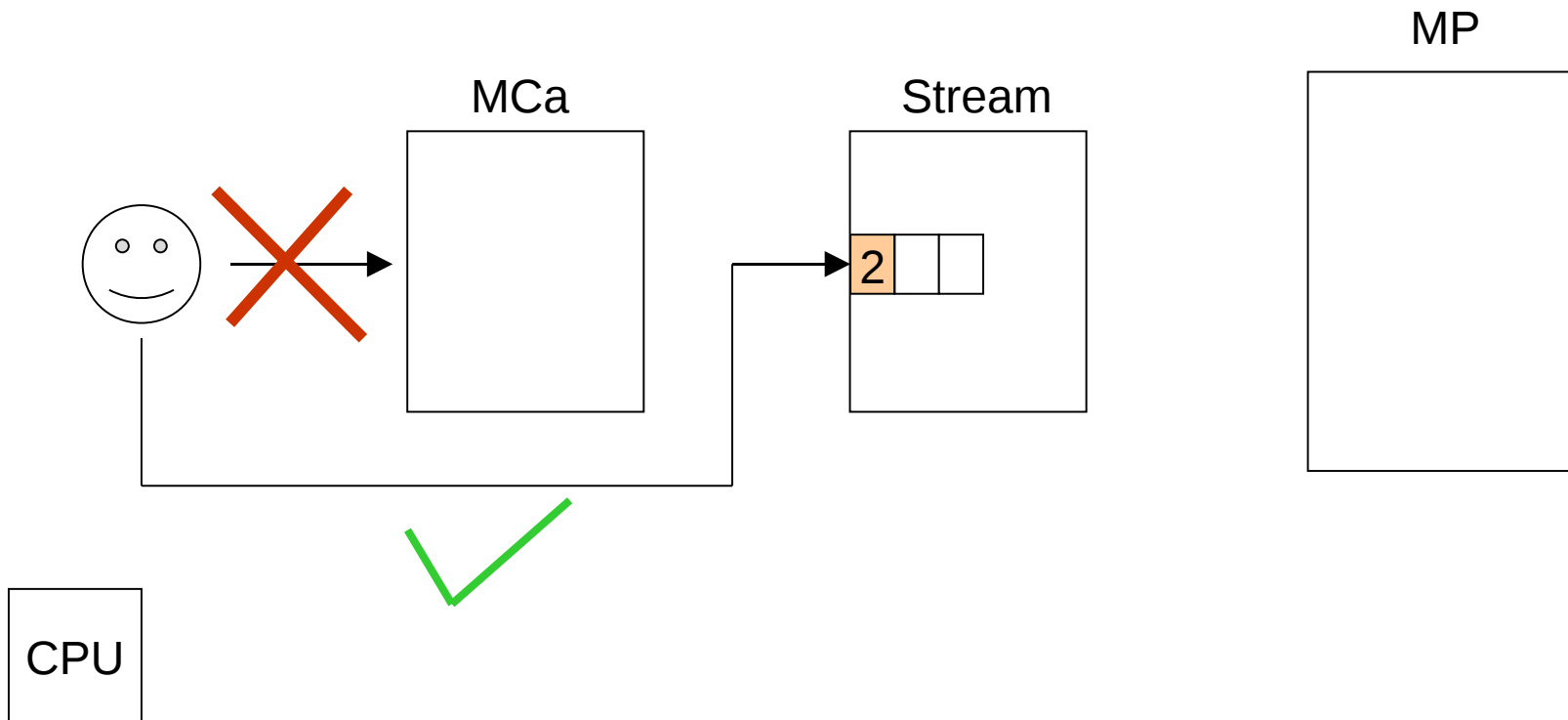
Stream Buffer

■ Anticipaciones



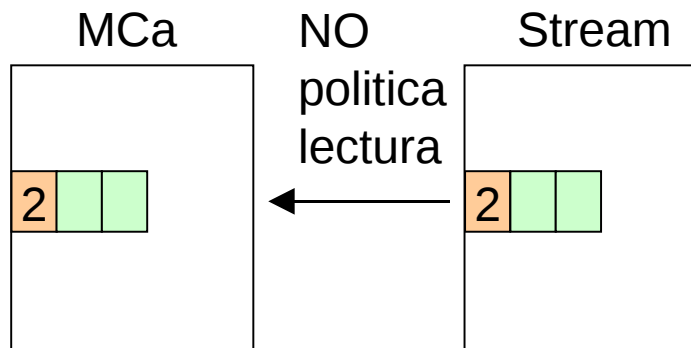
Stream Buffer

■ Anticipaciones

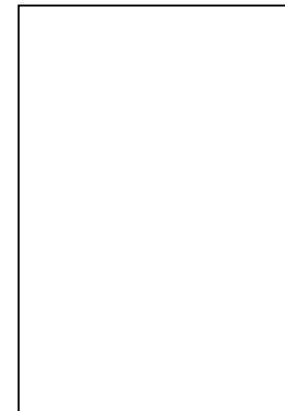


Stream Buffer

■ Anticipaciones



MP



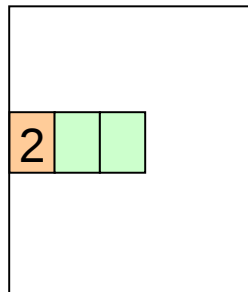
CPU

Stream Buffer

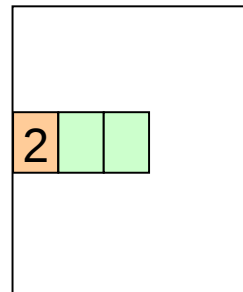
■ Anticipaciones



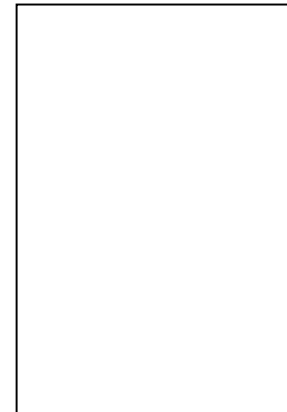
MCa



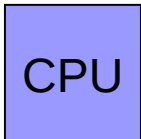
Stream



MP

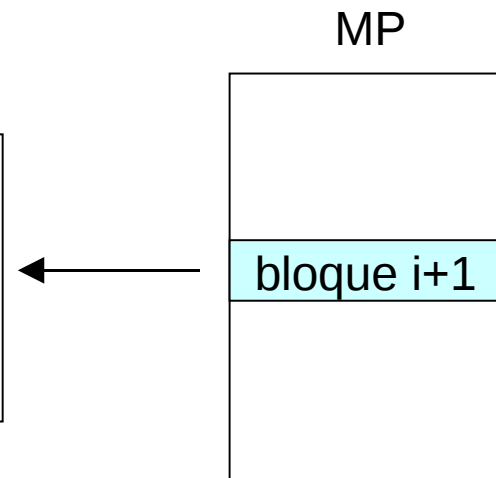
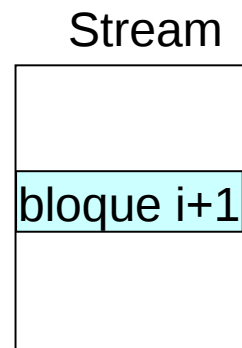
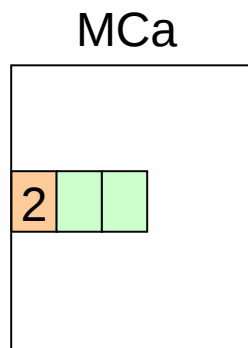


CPU



Stream Buffer

■ Anticipaciones



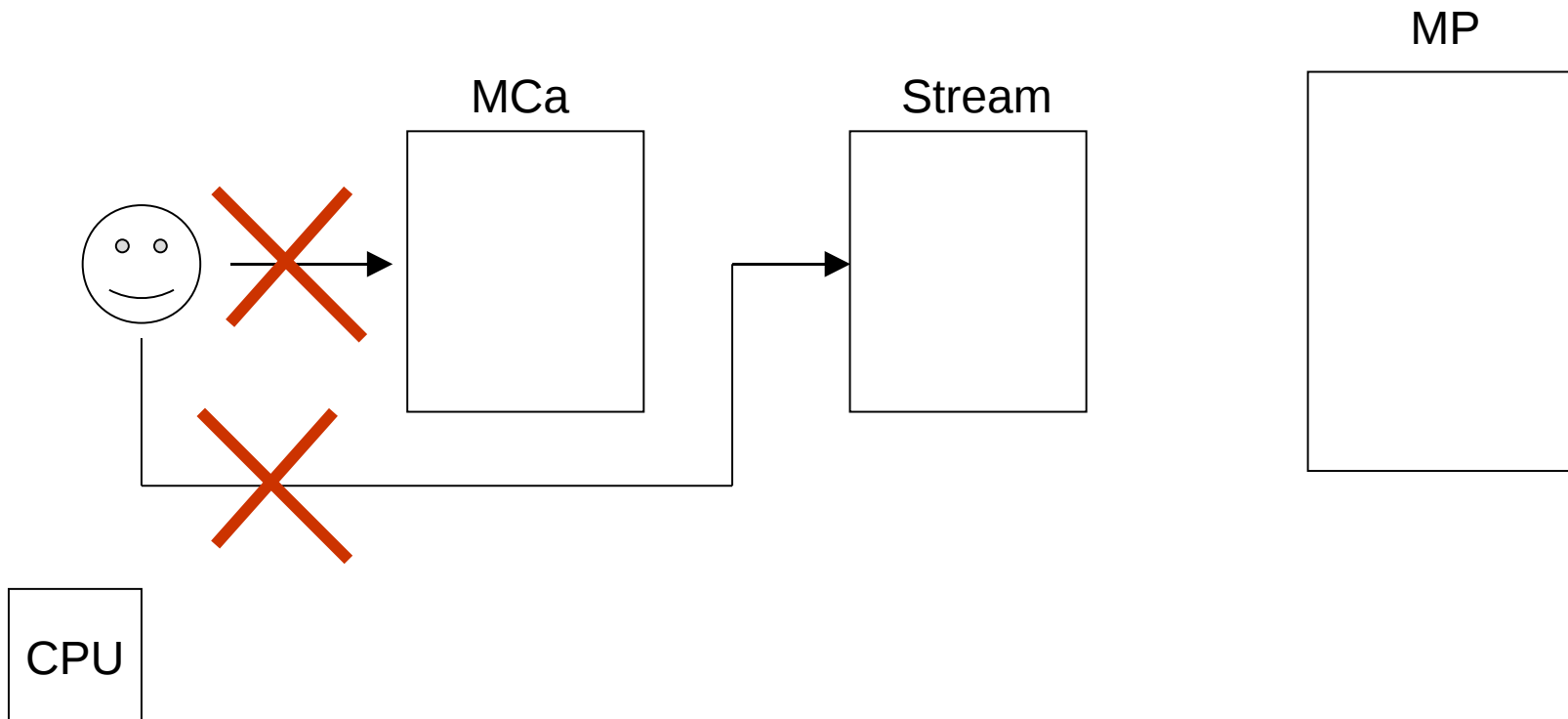
CPU

$$T_{acc} = T_{ca} + N * \max(MCa, Stream)$$

$$T_{ocup} = T_{ca} + N * \max(Mca, Stream) + N * T_{mp}$$

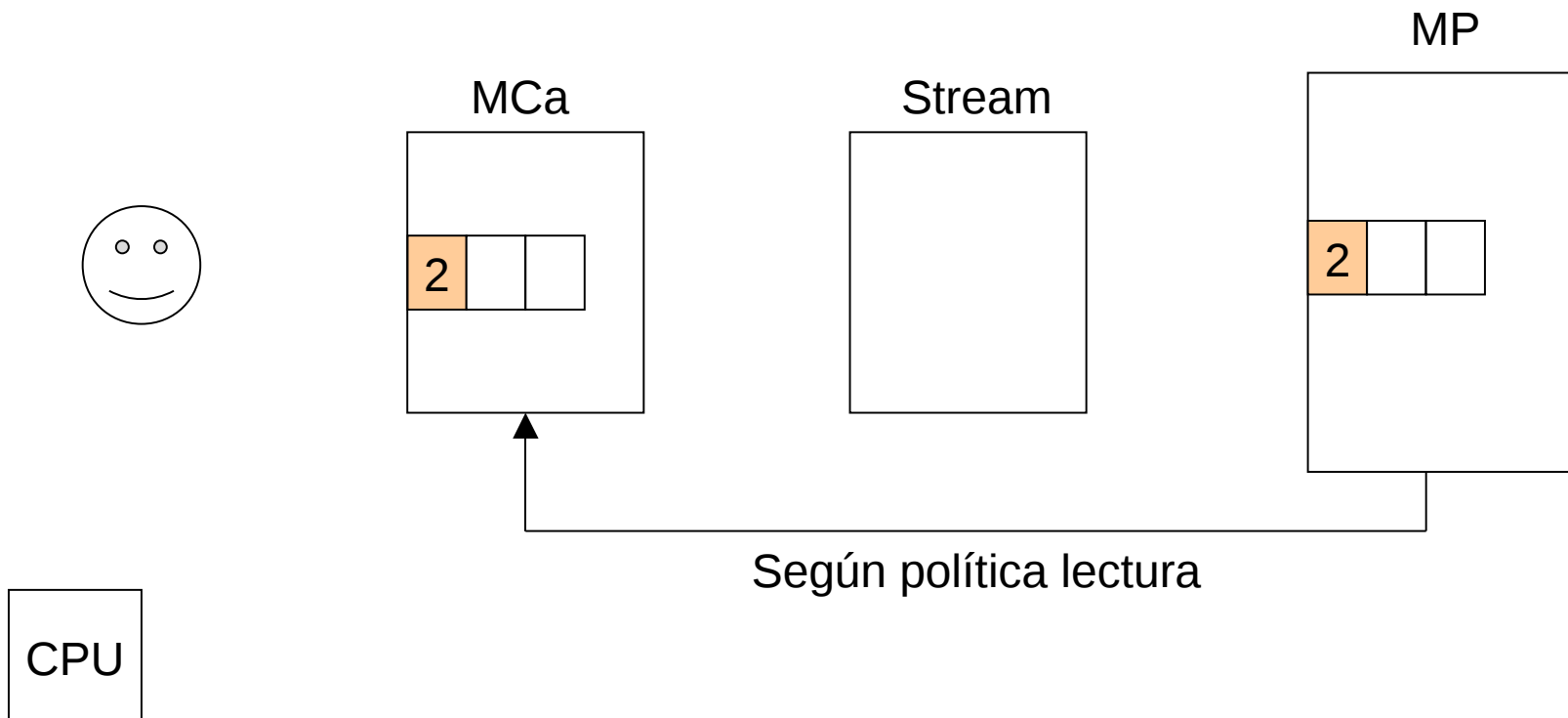
Stream Buffer

■ Anticipaciones



Stream Buffer

■ Anticipaciones

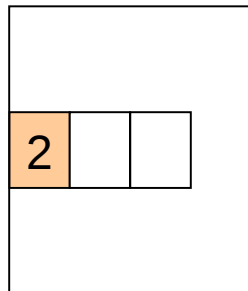


Stream Buffer

■ Anticipaciones



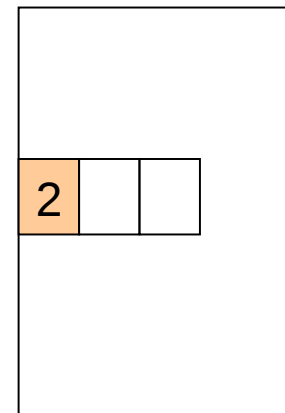
MCa



Stream



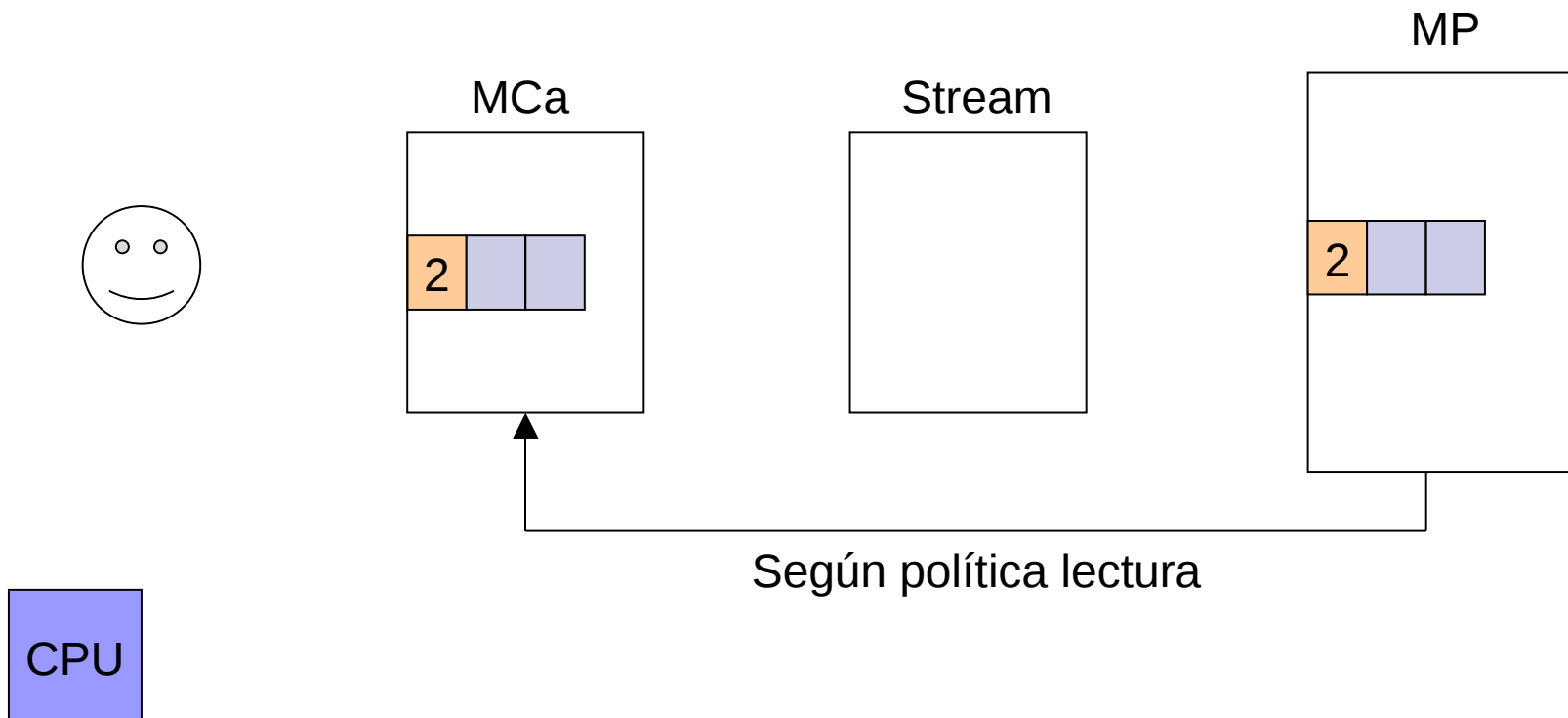
MP



CPU

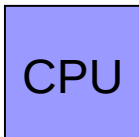
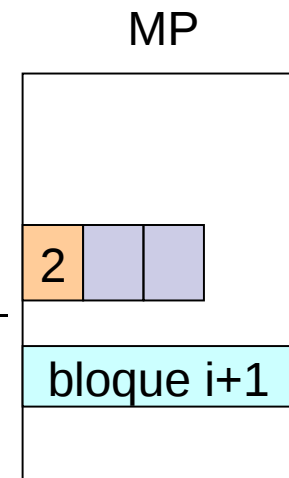
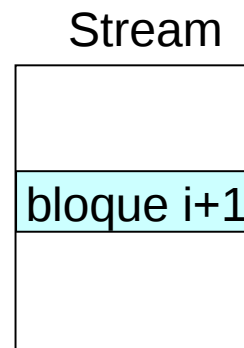
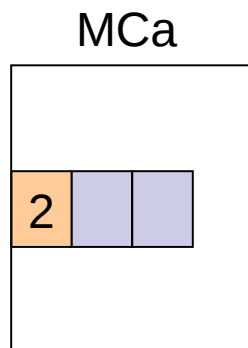
Stream Buffer

■ Anticipaciones



Stream Buffer

■ Anticipaciones



$$T_{acc} = T_{ca} + T_{str} + X * T_{mp}$$
$$T_{ocup} = T_{ca} + T_{str} + N * T_{mp} + N * T_{mp}$$

Buffer de escritura

- Escrituras de MP **NO** en Tacc (Tocup)
- Se utiliza este mecanismo **sii** buffer no lleno
- Probabilidades llenado de Buffer
 - $Pb_vacio * Tb_vacio + (1 - Pb_vacio) * Tacc_n$
 - $Tb_vacio = Tacc$ usando buffer de escritura



Buffer de escritura (II)

■ Casos de estudio

□ Política Escritura: Copy-Back

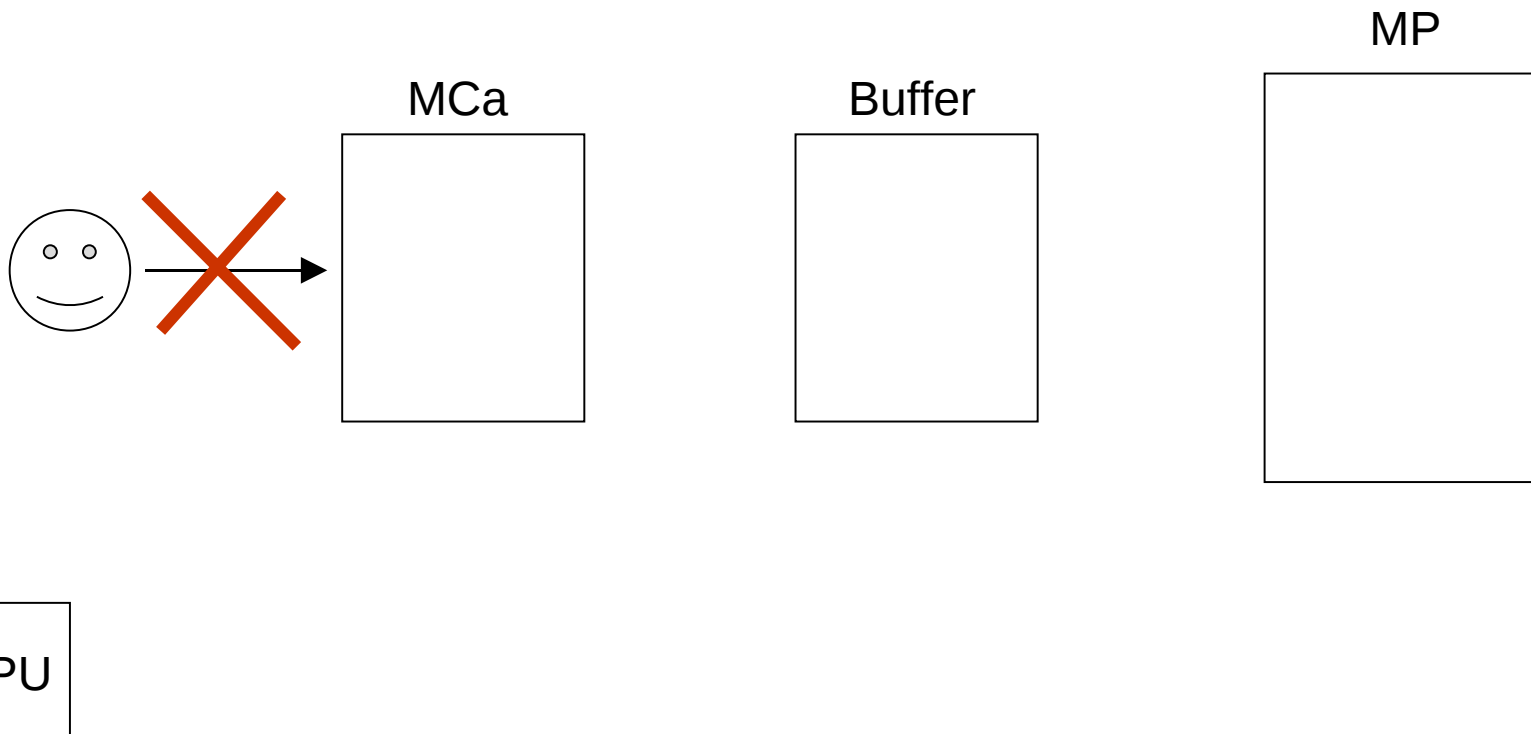
- Op Lectura, fallo MCA con reemplazo
- Op Escritura, fallo MCA con reemplazo

□ Política Escritura: Write-through

- Op Escritura, acierto o fallo

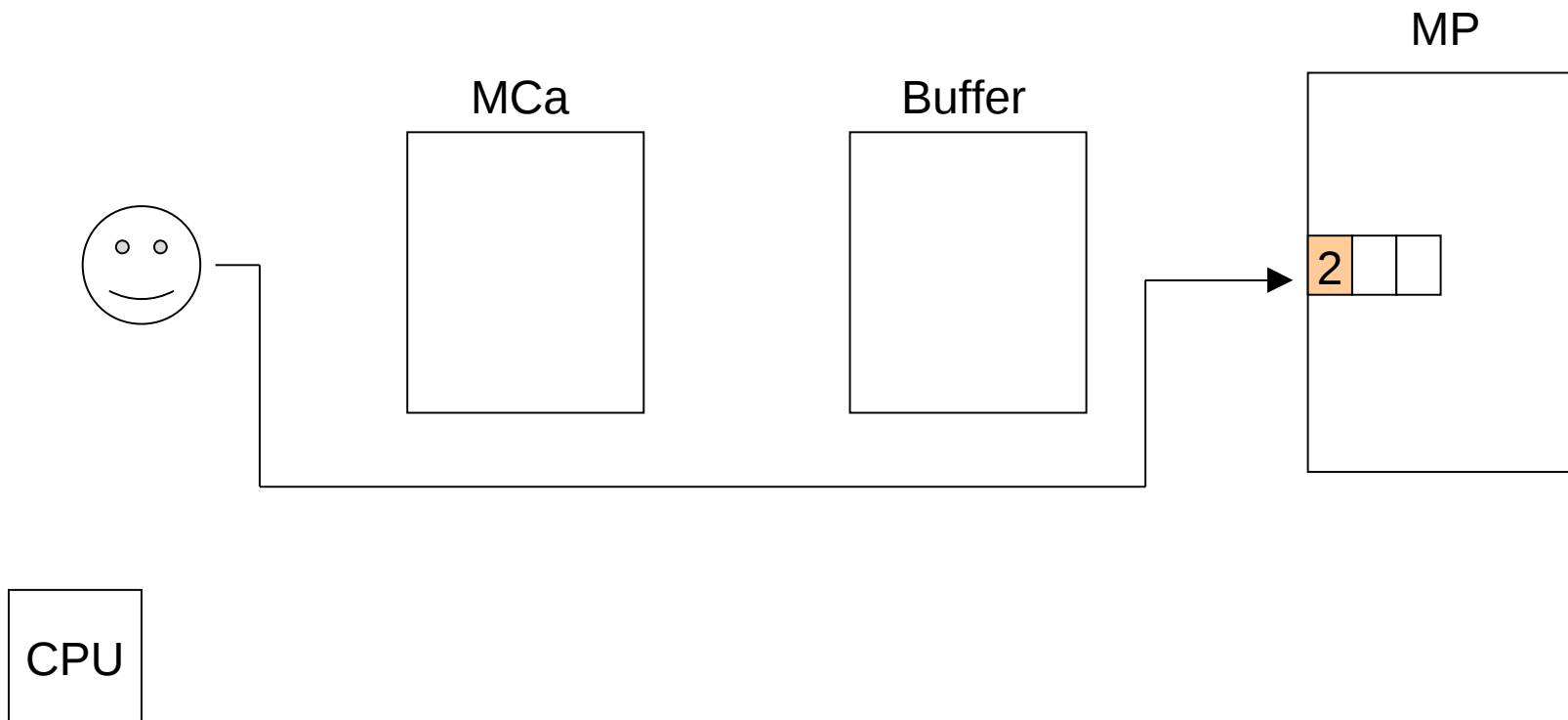
Buffer de escritura (III)

CB, Op lectura y escritura



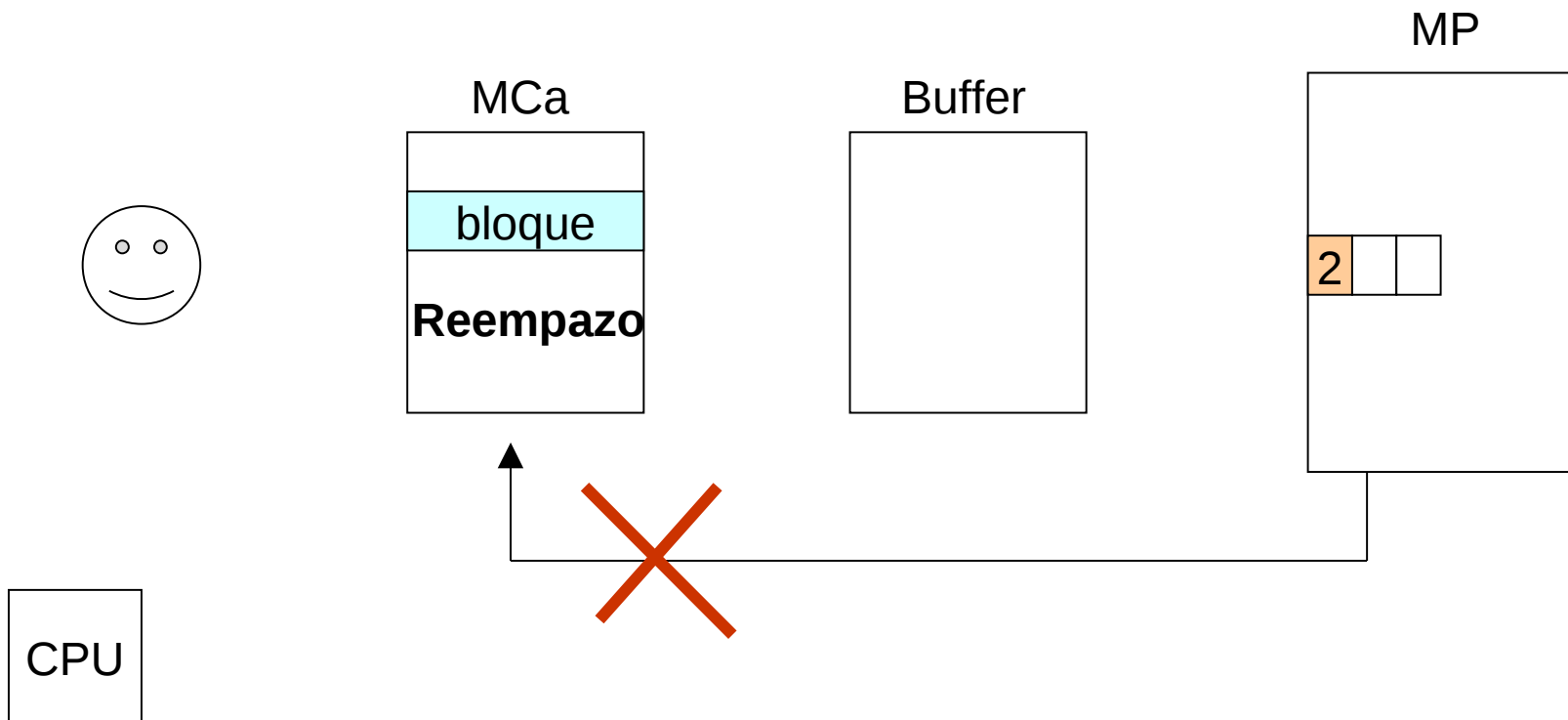
Buffer de escritura (III)

CB, Op lectura y escritura



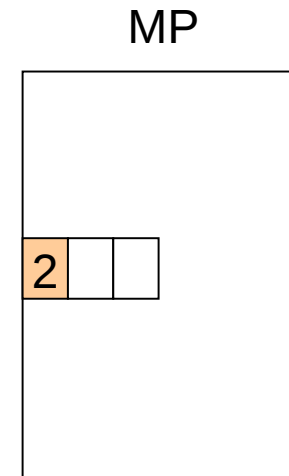
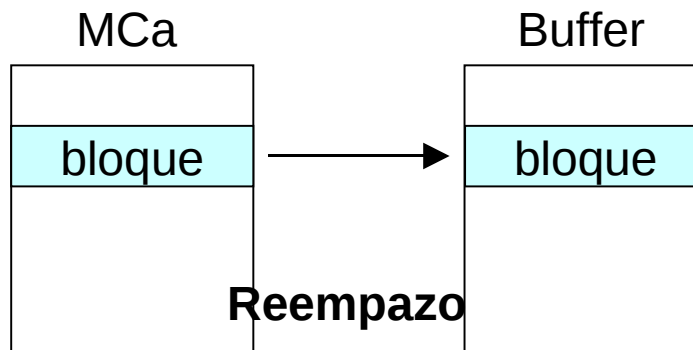
Buffer de escritura (III)

CB, Op lectura y escritura



Buffer de escritura (III)

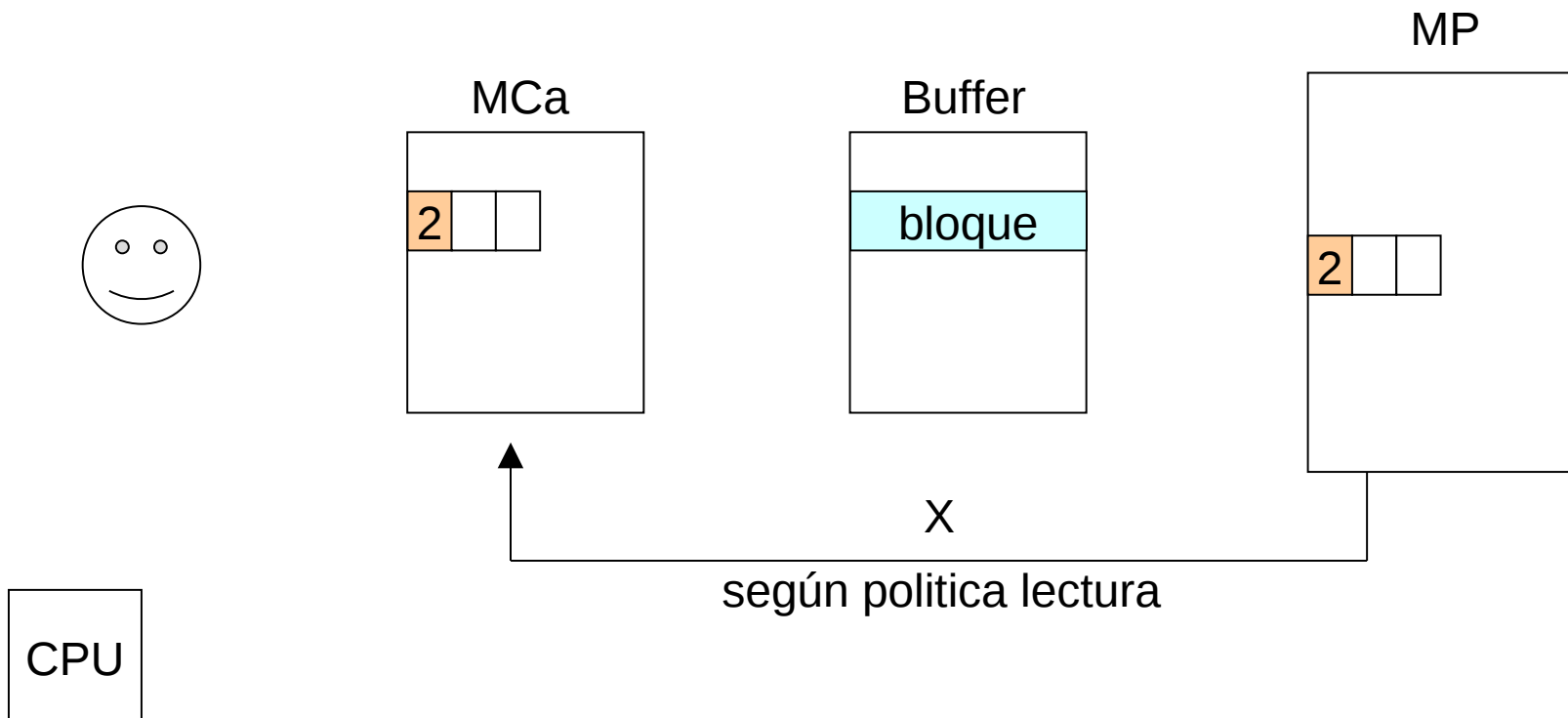
CB, Op lectura y escritura



CPU

Buffer de escritura (III)

CB, Op lectura y escritura

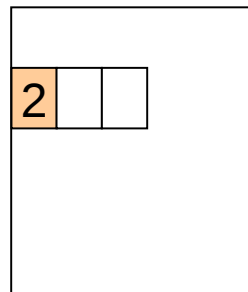


Buffer de escritura (III)

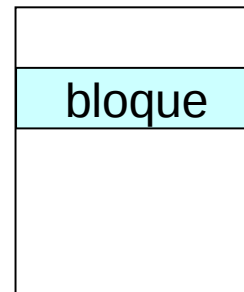
CB, Op lectura y escritura



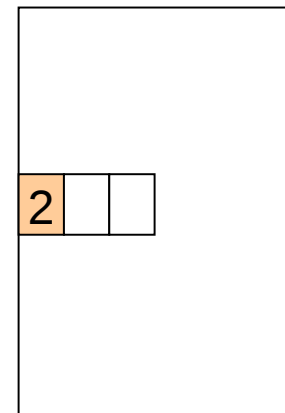
MCa



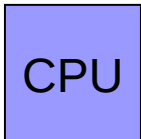
Buffer



MP

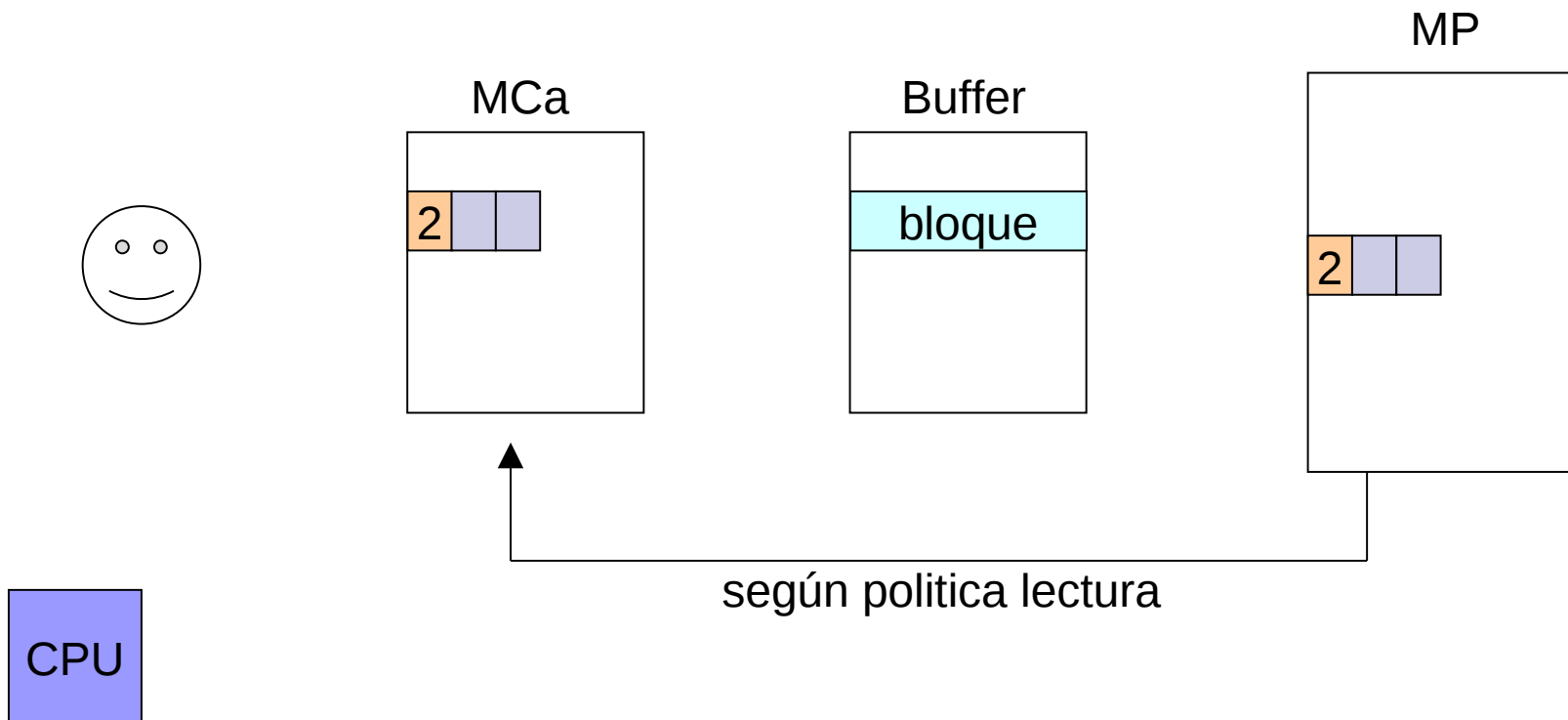


CPU



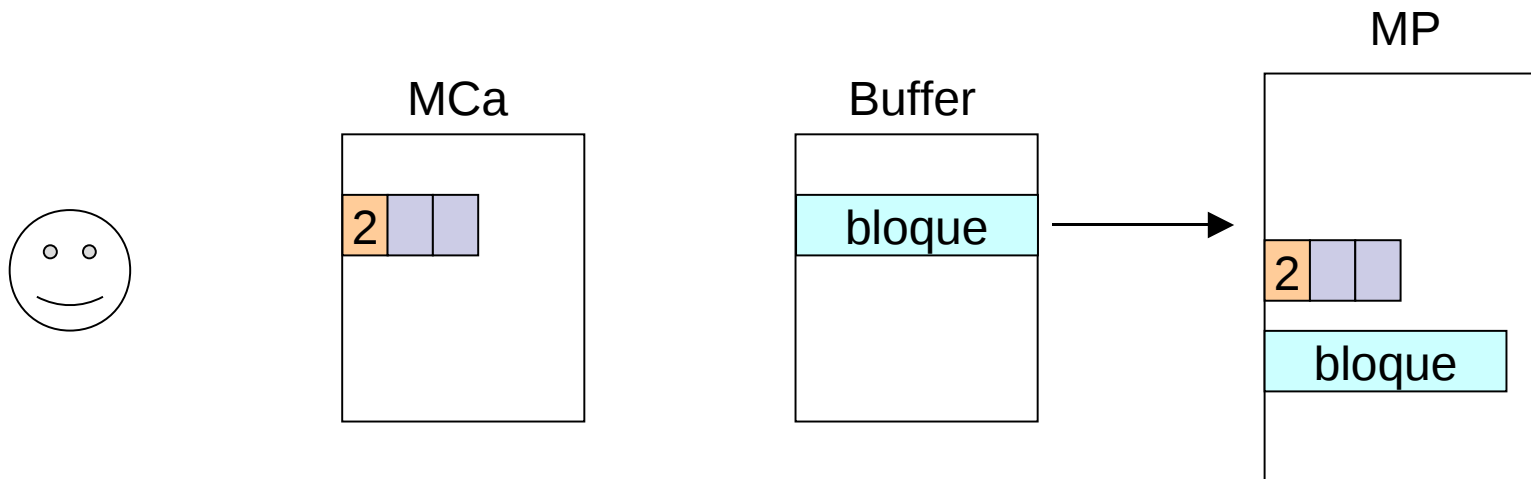
Buffer de escritura (III)

CB, Op lectura y escritura



Buffer de escritura (III)

CB, Op lectura y escritura



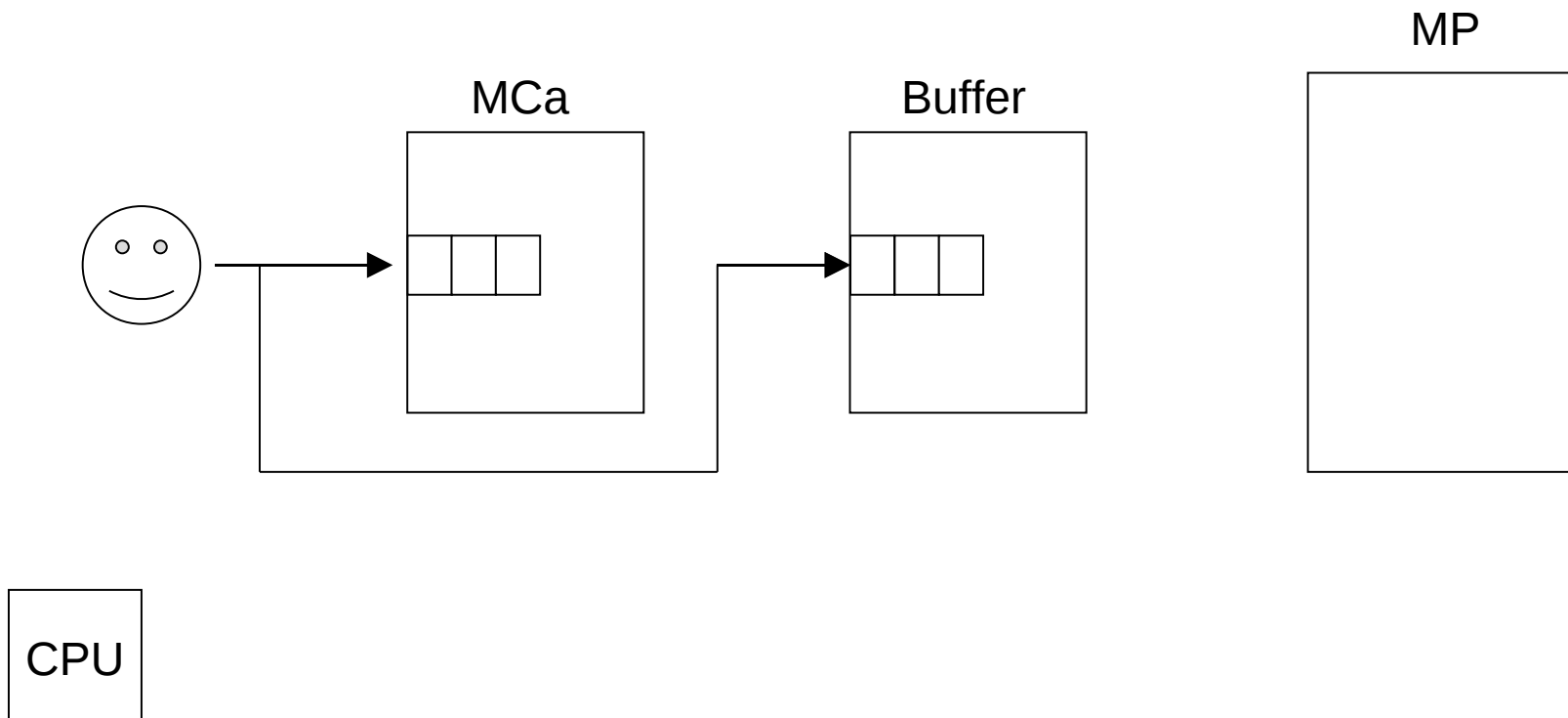
$$T_{acc} = T_{ca} + N * \max(T_{ca}, T_{buff}) + X * T_{mp}$$

$$T_{ocup} = T_{ca} + N * \max(T_{ca}, T_{buff}) + N * T_{mp} + N * T_{mp}$$

CPU

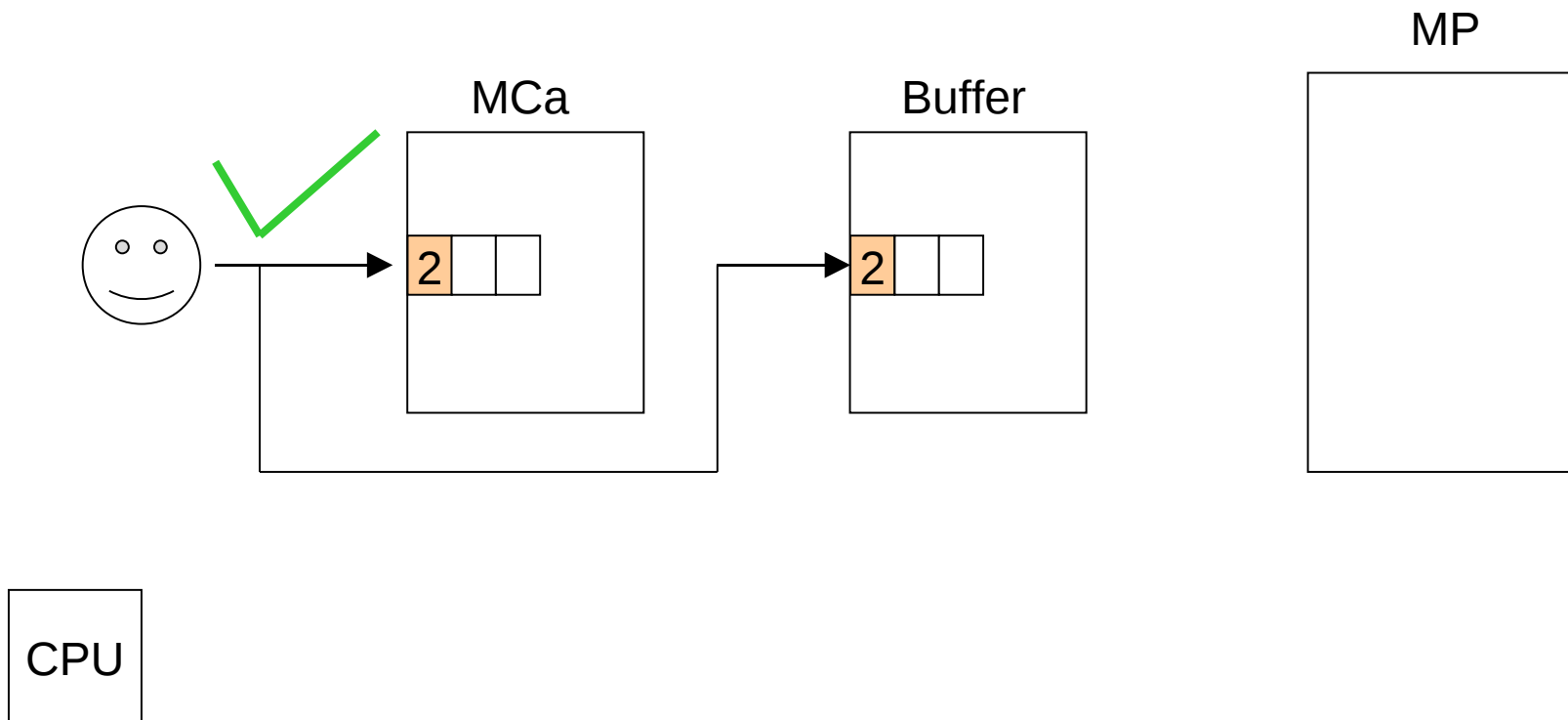
Buffer de escritura (III)

WT, Op escritura



Buffer de escritura (III)

WT, Op escritura

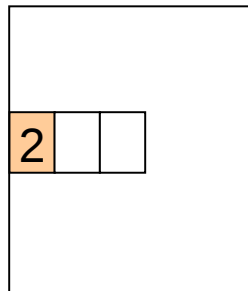


Buffer de escritura (III)

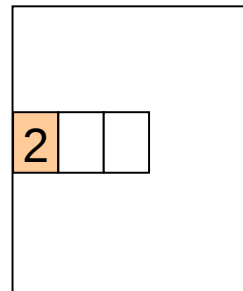
WT, Op escritura



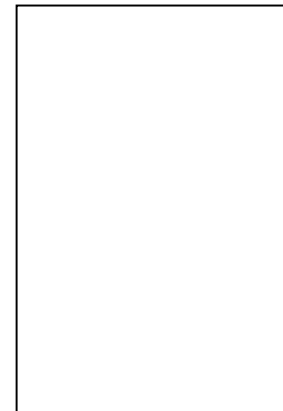
MCa



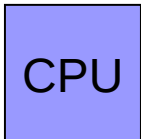
Buffer



MP



CPU

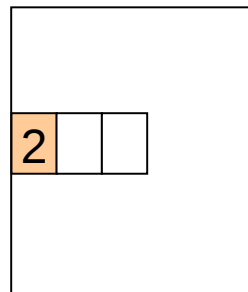


Buffer de escritura (III)

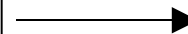
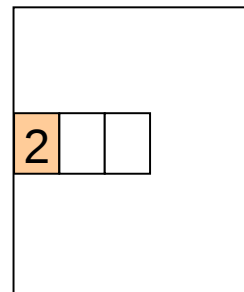
WT, Op escritura



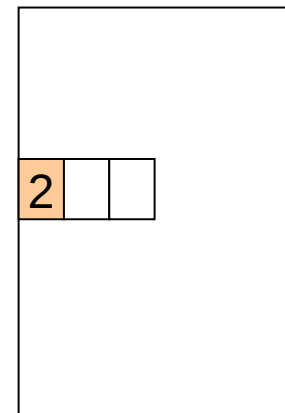
MCa



Buffer



MP



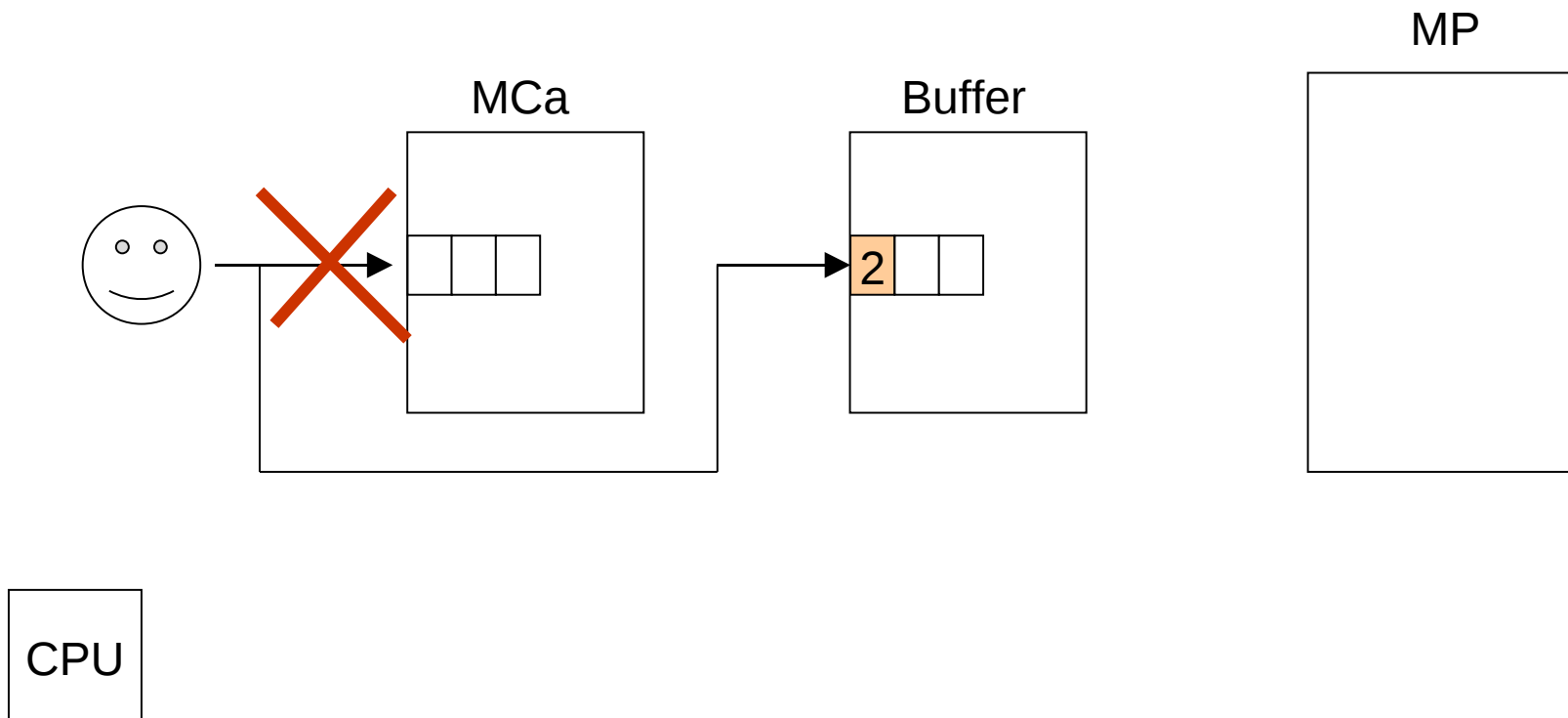
CPU

$T_{acc} = \max(T_{ca}, T_{buf})$

$T_{ocup} = \max(T_{ca}, T_{buf}) + T_{mp}$

Buffer de escritura (III)

WT, Op escritura

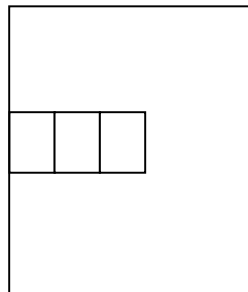


Buffer de escritura (III)

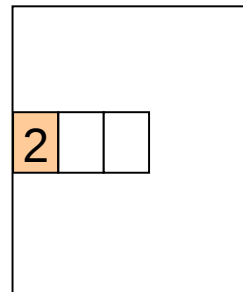
WT, Op escritura



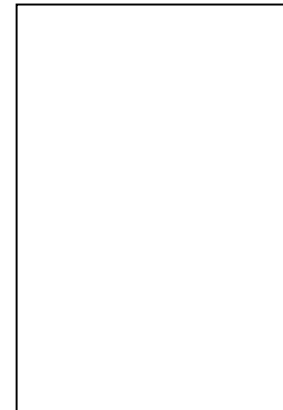
MCa



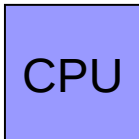
Buffer



MP



CPU

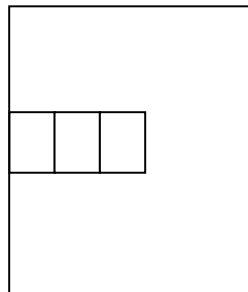


Buffer de escritura (III)

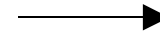
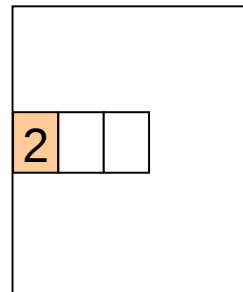
WT, Op escritura



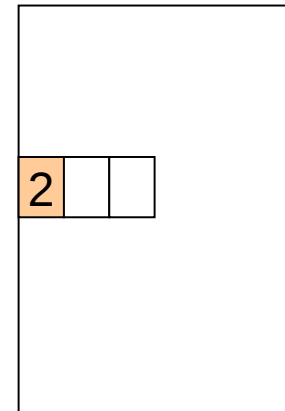
MCa



Buffer



MP



CPU

$$T_{acc} = T_{buf}$$

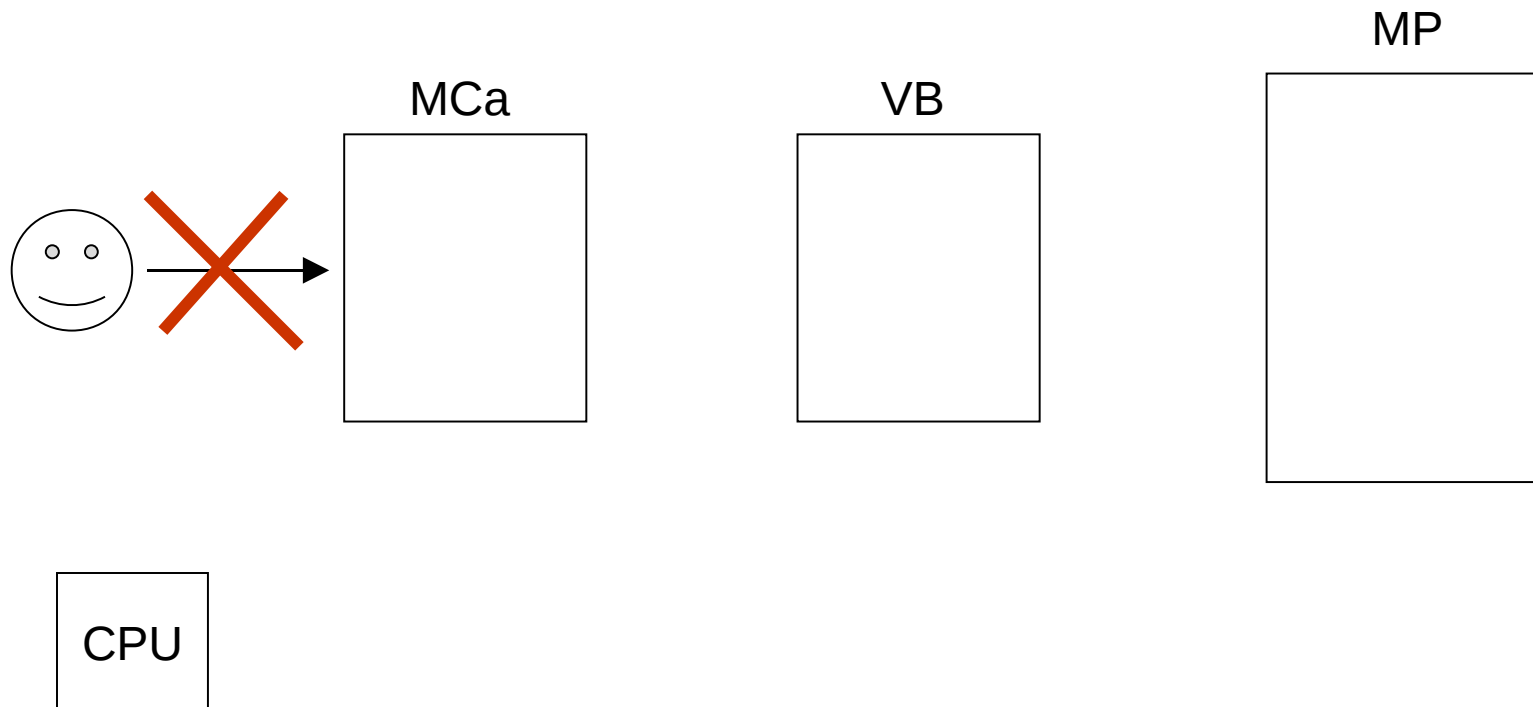
$$T_{ocup} = T_{buf} + T_{mp}$$



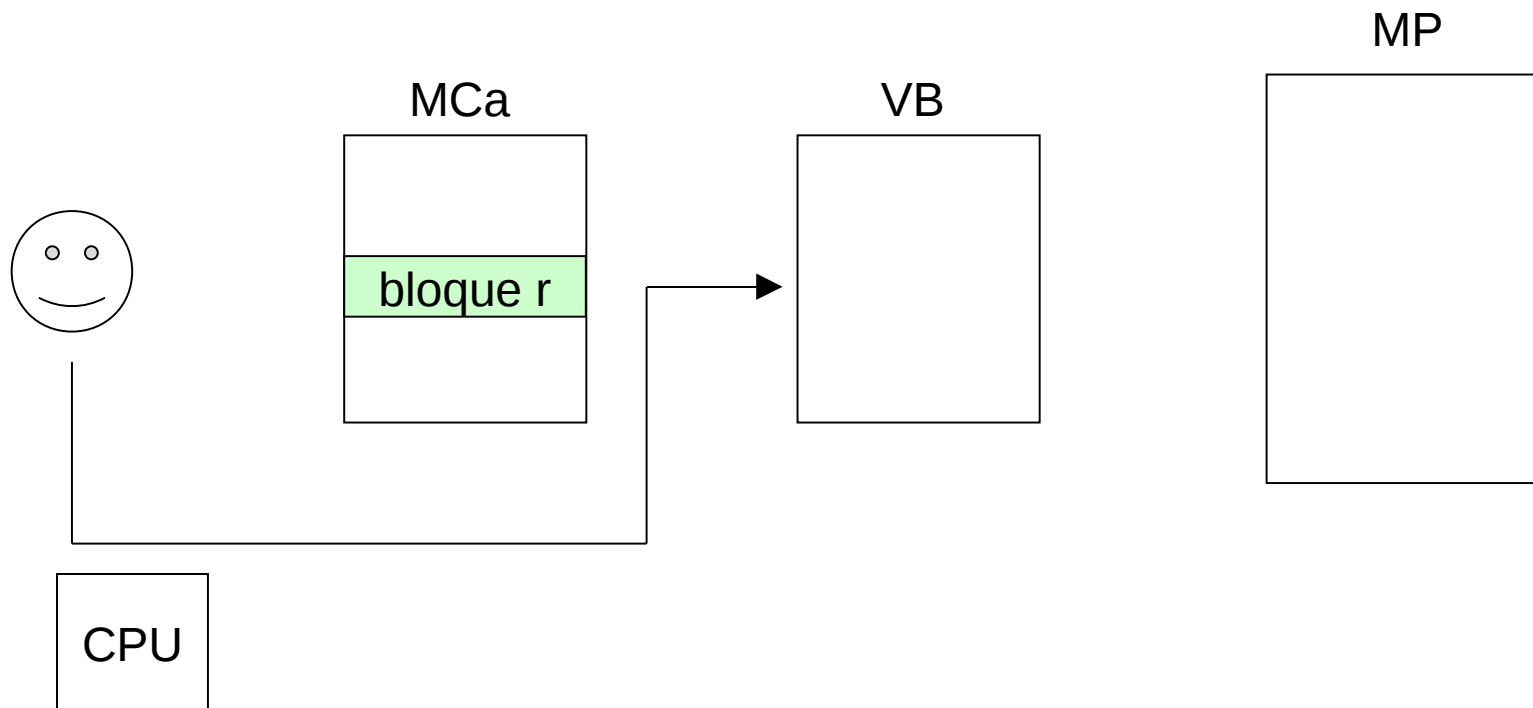
Victim Buffer

- Unicamente para **Caches DIRECTAS**
- Reducir el Tacc por **fallos de conflicto**
(dos palabras en la misma dir)

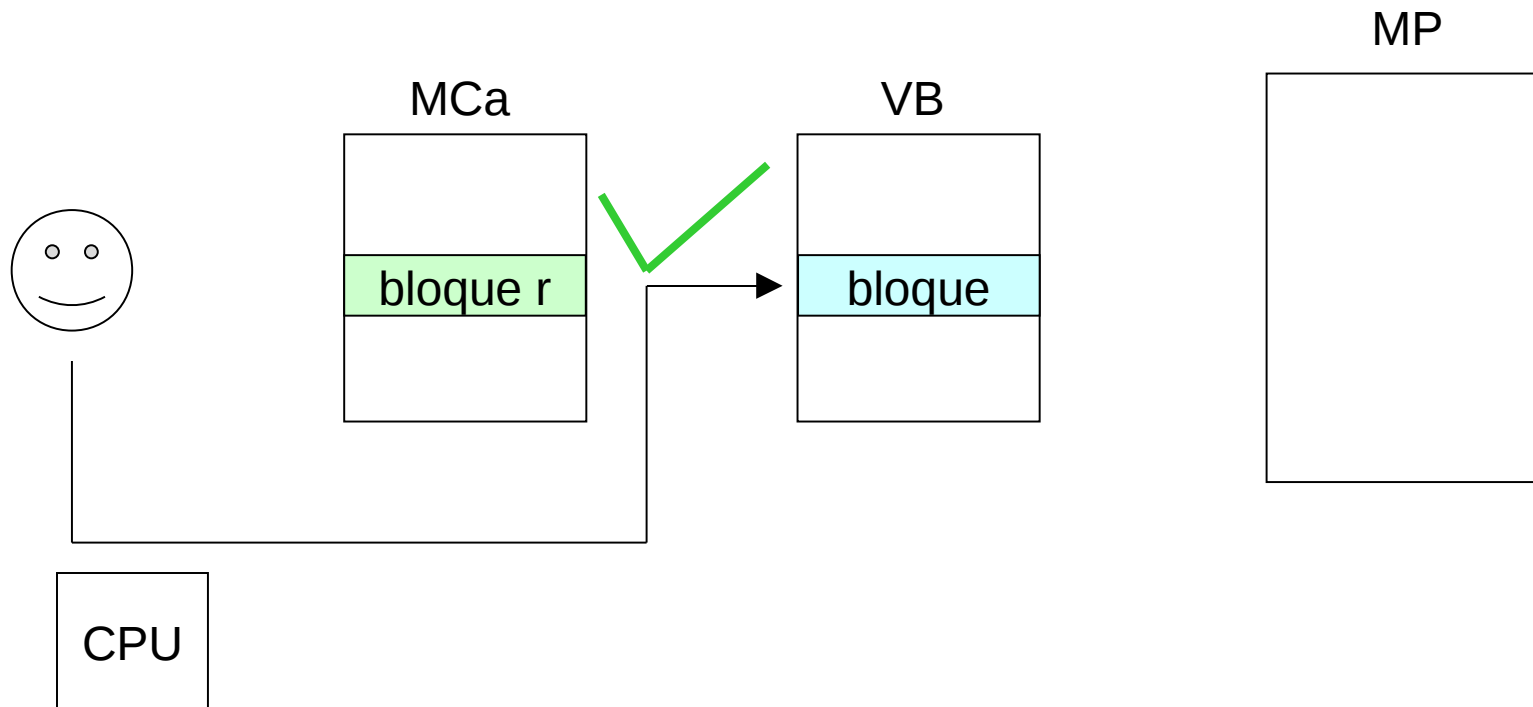
Victim Buffer (II)



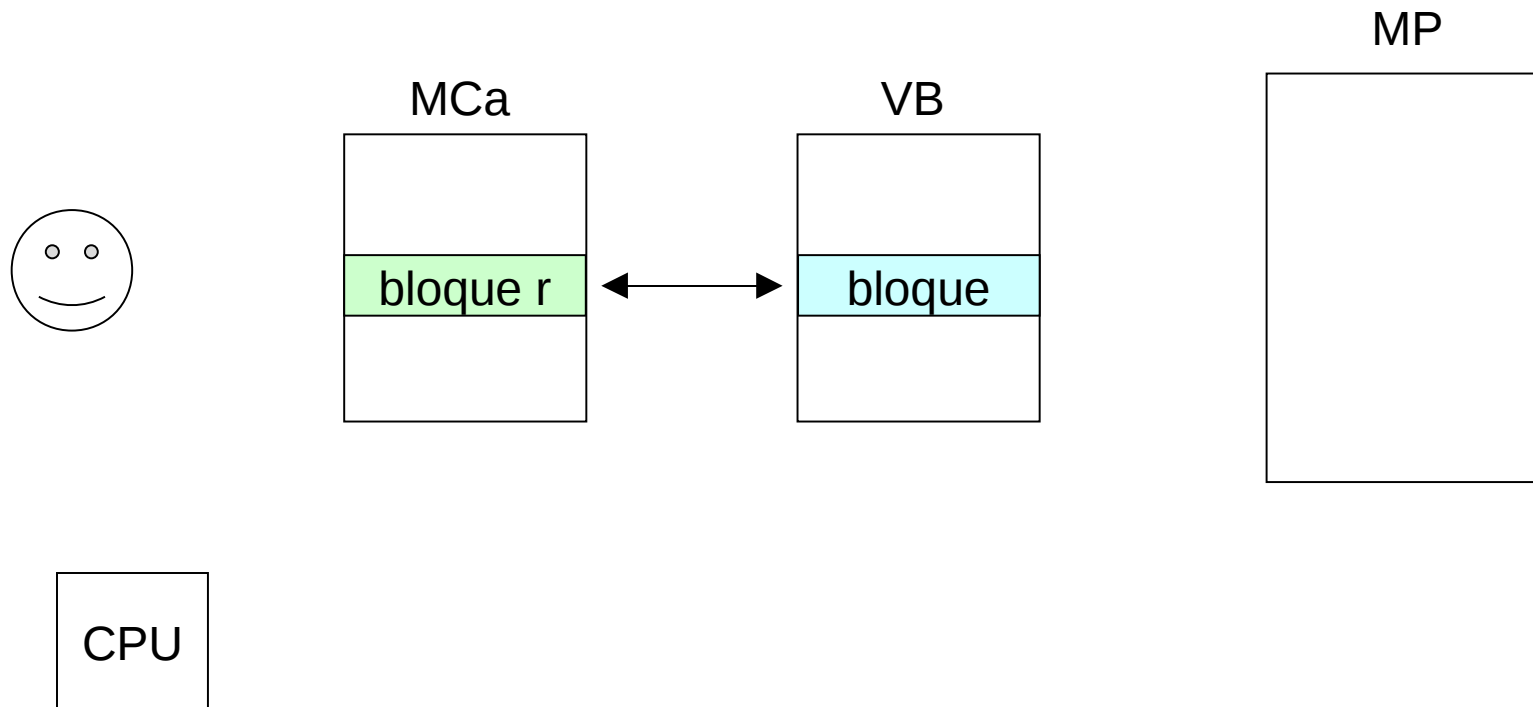
Victim Buffer (II)



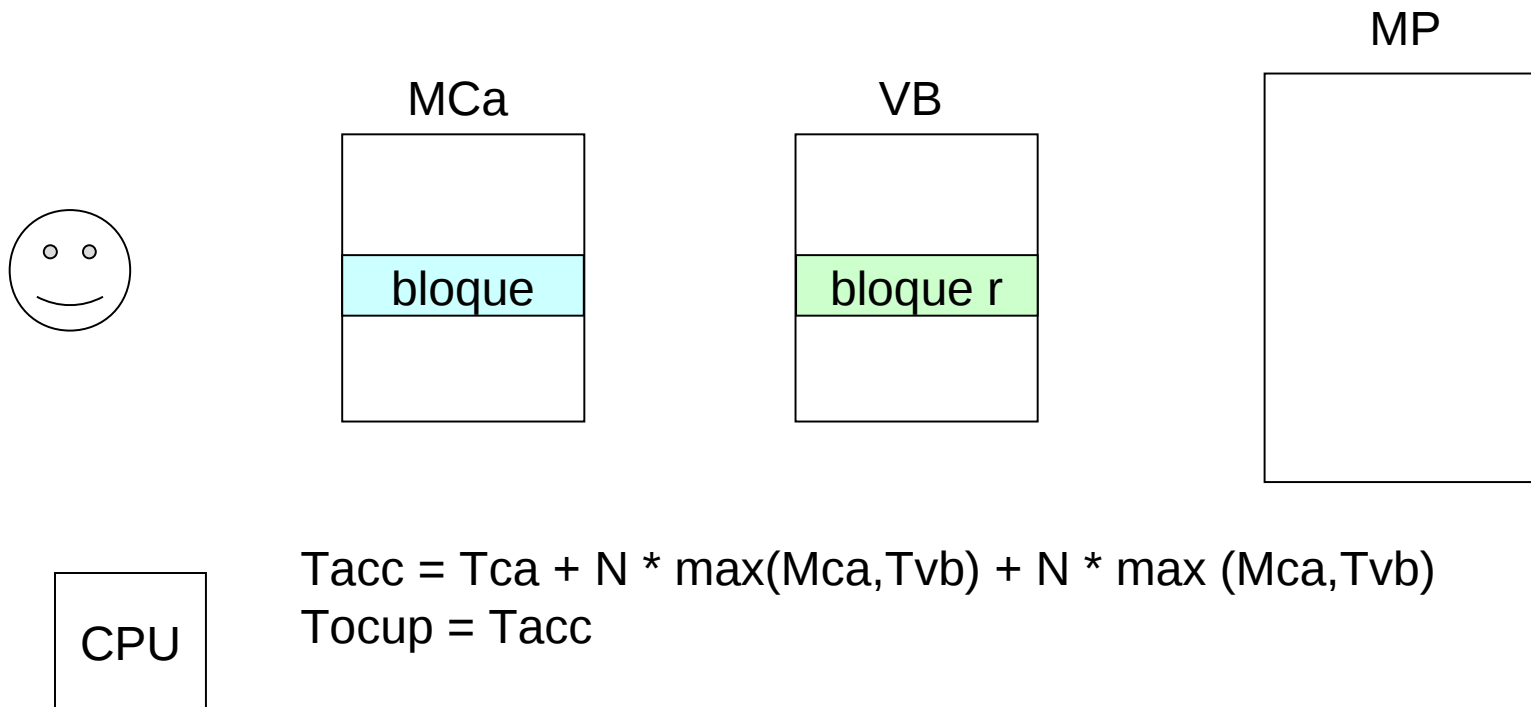
Victim Buffer (II)



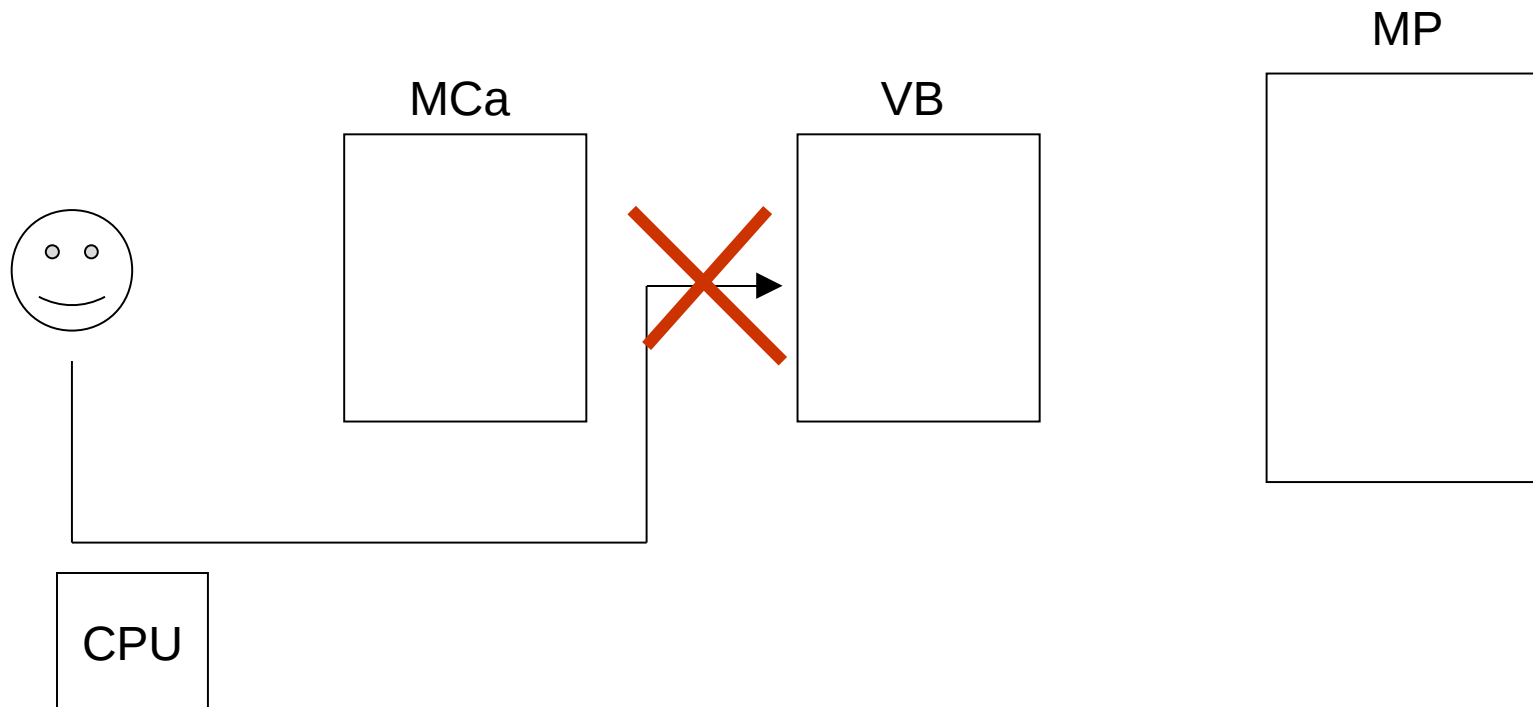
Victim Buffer (II)



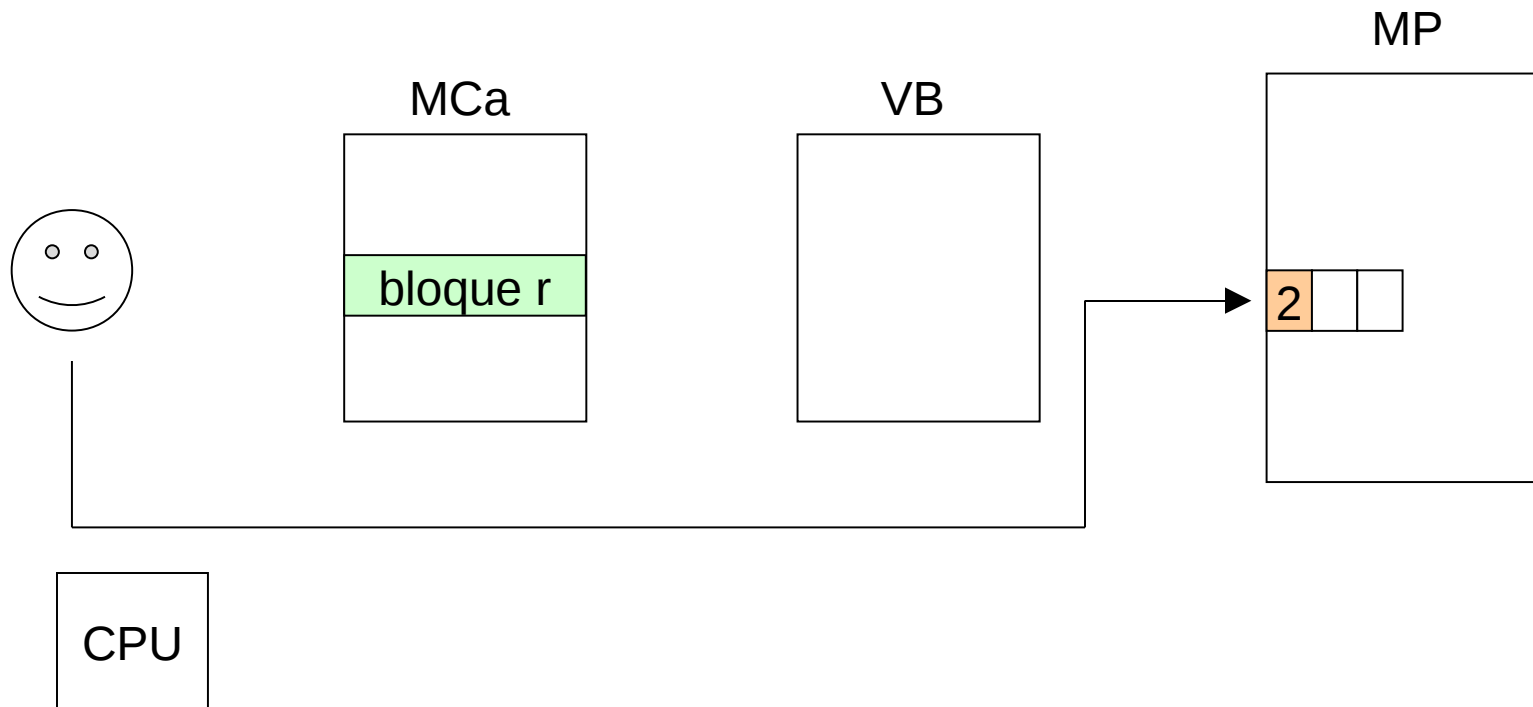
Victim Buffer (II)



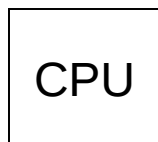
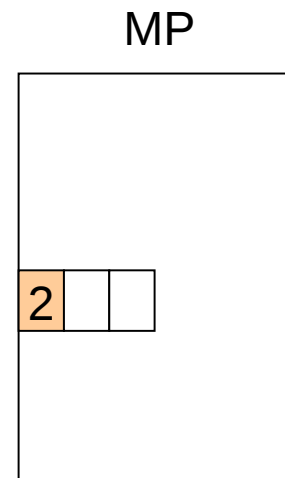
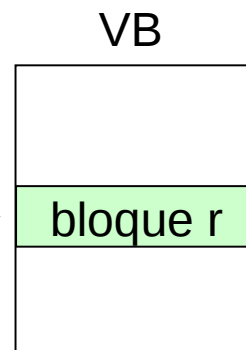
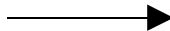
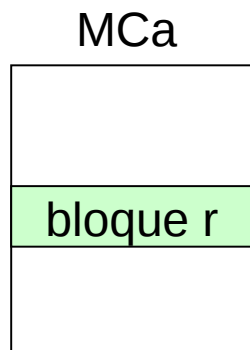
Victim Buffer (II)



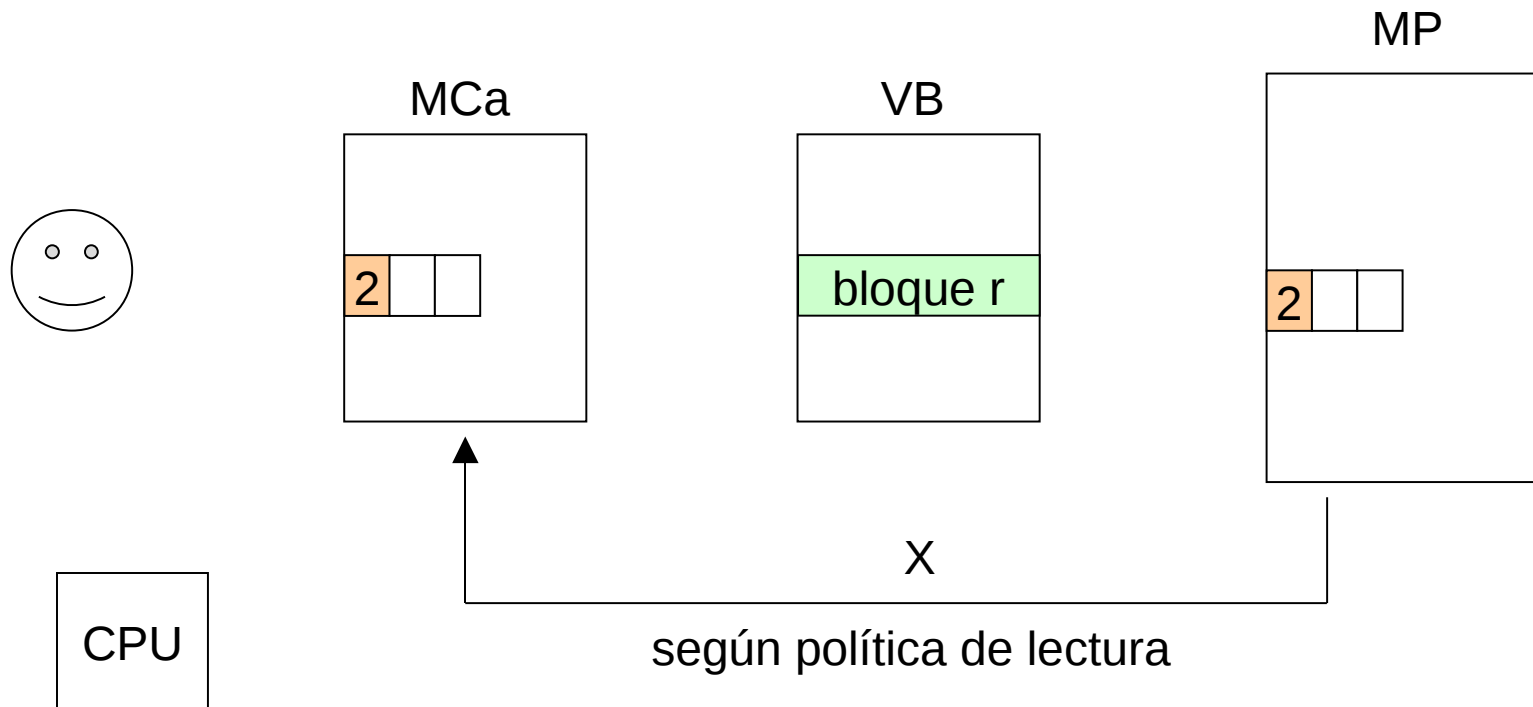
Victim Buffer (II)



Victim Buffer (II)



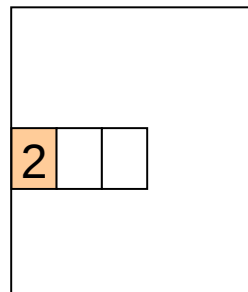
Victim Buffer (II)



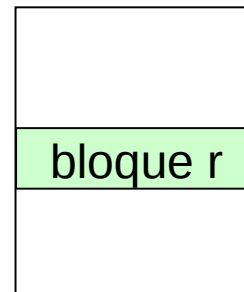
Victim Buffer (II)



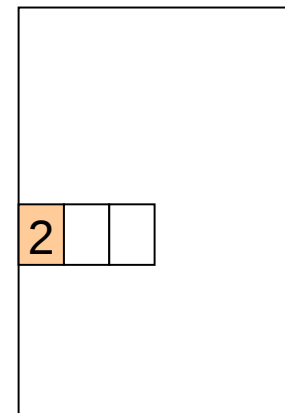
MCa



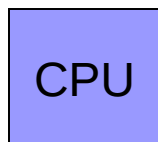
VB



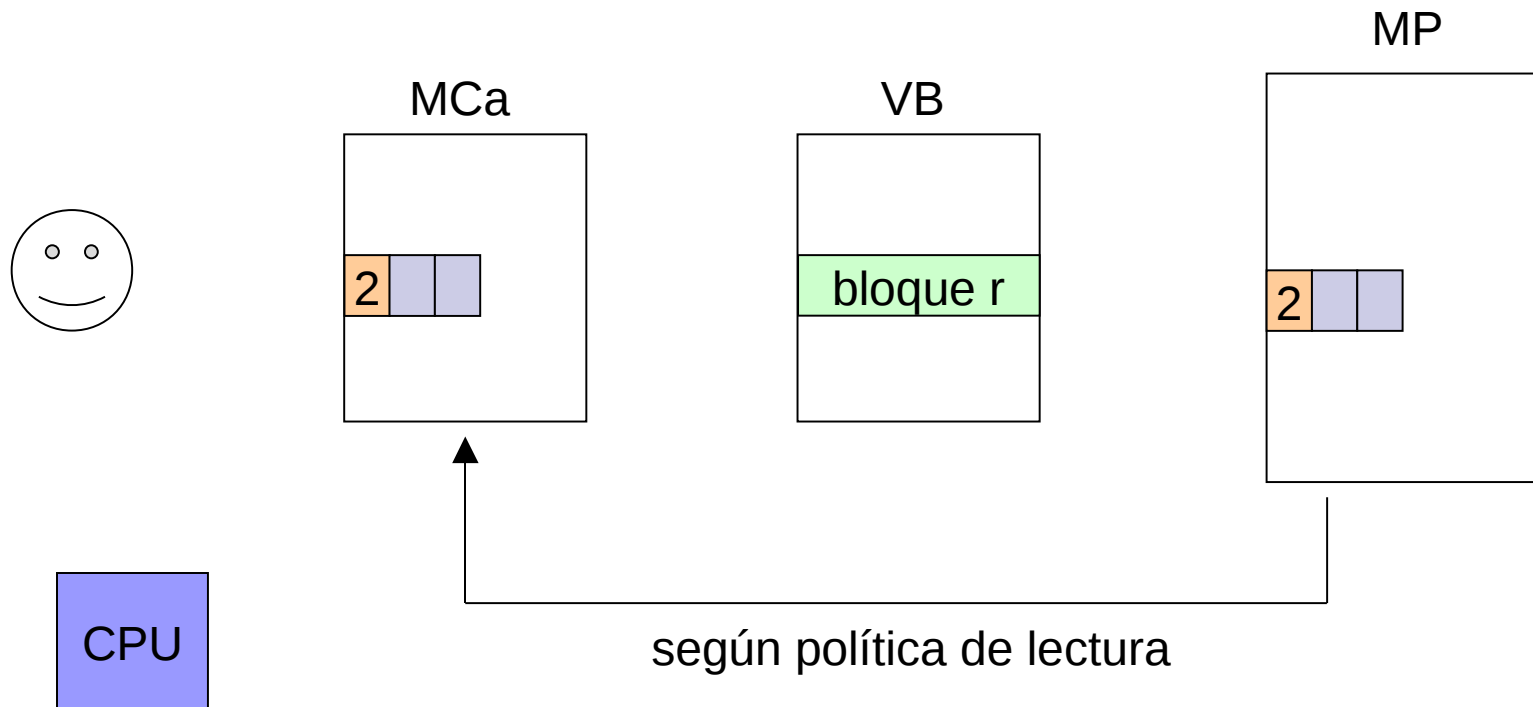
MP



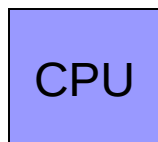
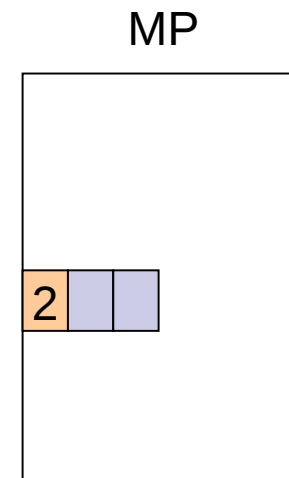
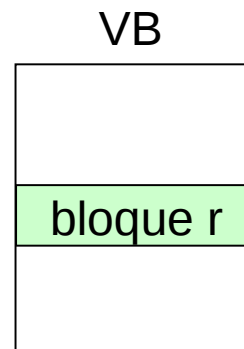
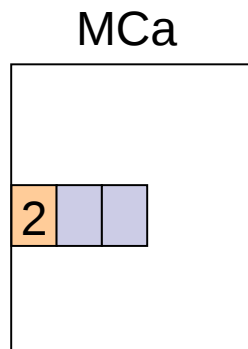
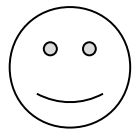
CPU



Victim Buffer (II)



Victim Buffer (II)



CPU

$$T_{acc} = T_{ca} + T_{vb} + N * \max(T_{ca}, T_{vb}) + X * T_{mp}$$

$$T_{ocup} = T_{ca} + T_{vb} + N * \max(T_{ca}, T_{vb}) + N * T_{mp}$$



Memorias Cache

Fin